Situational Analysis of the Sacrifice Bunt in Baseball

by

Mathew Chandler

A Thesis

Submitted to the Faculty of Graduate Studies and Research

through the Department of Mathematics and Statistics

in Partial Fulfillment of the Requirements for

the Degree of Master of Science at the

University of Windsor

Windsor, Ontario, Canada

1997

0-612-30937-1

ABSTRACT

The strategy of the sacrifice bunt is a situational strategy which is most often used when the scoring of a single run is considered to be an important objective. Despite this, existing research has focused on determining whether the sacrifice bunt is useful in a general context. The objective of this thesis is to attempt to develop a method to determine the strategy giving the higher probability of scoring at least one run-attempting a sacrifice bunt or batting normally (when the situational variables are essentially neutral and do not suggest a course of action). This will be done by simulating a large number of half innings in which each of the two options is used. The output of the simulations will then be used to perform two regression analyses, which will generate two equations. The first equation will equate the probability of scoring at least one run to some combination of sacrifice bunting efficiency, team batting average and team slugging percentage. The second equation will equate the probability of scoring at least one run to some combination of the potential bunter's slugging percentage, team batting average and team slugging percentage. The two equations will then be subtracted and simplified to produce one equation that equates

sacrifice bunting efficiency to some combination of the potential bunter's slugging percentage, team slugging percentage and team batting average. This will allow us to compare the two strategies in the form of a single equation. This is done by simply entering the values of the potential bunter's slugging percentage, the team slugging percentage and the team batting average in the equation. This will generate the sacrifice bunting efficiency that has an equal probability to score at least one run in the given situation. Therefore, if the actual sacrifice bunting efficiency is greater than this generated value, then the better strategy is to sacrifice bunt. The strategies can also be examined graphically, where the graph of the equation generated represent where the two probabilities involved are approximately equal. Therefore, the better strategy can be determined based on which side of the curve the values of the variables occur. This approach to decision analysis, to our knowledge, has not been used before. This thesis will also consider the advantages of this method, examine some of the different considerations involved in whether this method can be used in other similar situations and try to specify what is necessary to be able to apply this method to other types of decision analysis problems.

ACKNOWLEDGEMENTS

I would like to acknowledge the financial support of the University of Windsor during the research and writing of this thesis. I would also like to acknowledge the advice and help of my advisor, Dr. Myron Hlynka, as well as Dr. Richard Caron.

TABLE OF CONTENTS

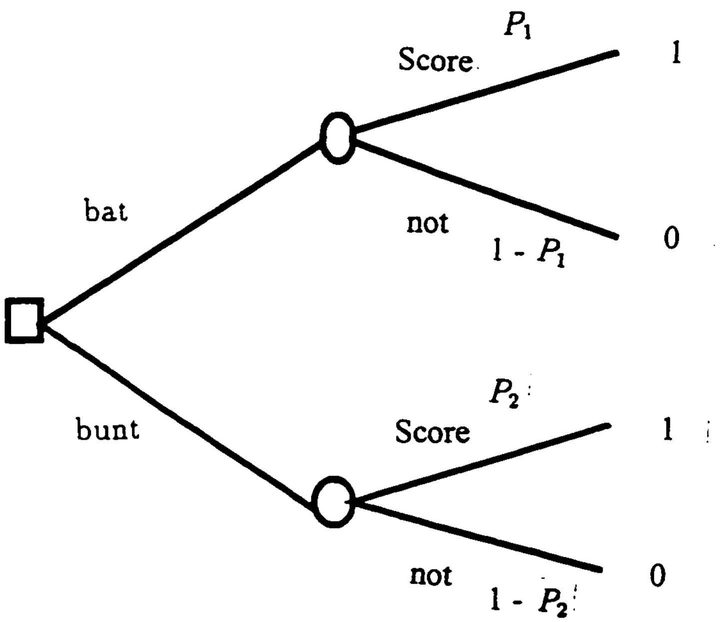
LIST OF ILLUSTRATIONS

1 Introduction

There have been many articles written about various aspects of baseball. There are many reasons for this. First, there have been extensive records kept for many decades of all Major League Baseball games, which provides a vast resource of data for research concerning baseball. Second, unlike most other professional sports, the non-continuous nature of the action in baseball makes it relatively simple to classify and analyze. It is easy to classify all results of a single play and, because there can only be a finite number of plays in any one game, it is relatively simple to determine how often each result occurs. Therefore, just about every aspect of the game of baseball can be analyzed and studied. Other prominent sports, such as football, hockey, or basketball, are not suited to such broad analysis. In football, you can analyze whether or not to try a long field goal, depending on the field position of the team, or in hockey, where you can study whether or not to remove the goaltender and how early to remove the goaltender, when trailing by one goal late in a game. Analysis of player value can be done in just about any sport.

In many articles concerned with baseball strategy, the effectiveness of using the sacrifice bunt has been considered. In these articles, the authors have focused on the general usefulness of the sacrifice bunt and usually relied on data from actual player performances and games played to support their conclusions. Examples are "A Scientific Approach to Strategy in Baseball" by G. Lindsey (1977) in which the games played during the 1958, 1959 and 1960 Major League baseball seasons were used as evidence for the effectiveness of the suggested strategies, and "Analysis of Baseball as a Markov Process" by R. Trueman (1977) which relied on the season averages of the most regularly used lineup players of the 1973 Los Angeles Dodgers as its source of data. It has been found that the sacrifice bunt is not a useful play, *in general.* It is usually acknowledged, however, that the sacrifice bunt could be considered useful in the late innings of games where the scoring of one run is an important objective (i.e. when the game is tied or the batting team is behind by one run). It should be noted that, except when the situation would strongly suggest attempting a sacrifice bunt, such as a pitcher hitting with less than two outs and a man on base, or when facing a pitcher who is pitching extremely well in a close game, the sacrifice bunt is only used in these late inning situations. Therefore, there is not really anything surprising about their results since the sacrifice bunt is not a strategy that is used, *in general.* Previous articles have not, however, attempted to determine in these late inning situations, when the manager might consider using a sacrifice bunt, when is it more probable to score with a sacrifice bunt rather than with the batter hitting normally or swinging away in the hope of scoring one run. The results of these previous articles are of limited applicability for trying to make this determination since, in general, the ability to sacrifice bunt has been ignored as a factor. Also, in most articles, the presence of any situational variables have been ignored.

Diagram1: Tree Diagram of Sacrifice Bunt Situation



In this situation, P_1 corresponds to the probability of scoring at least one run given that the batter bats normally, and P_2 corresponds to the probability of scoring at least one run given that the batter sacrifice bunts. In terms of expected reward, the batting normally option has an expected reward of P_1 and the expected reward of the bunting option is P_2.

The objective of this investigation will be, using a computer program to simulate a half inning of baseball, to develop a method to determine whether a sacrifice bunt is advisable in any situation where the manager considers the scoring of one run to be important. The method could be applied when the situational variables do not strongly suggest what would be the best strategy and would make use of certain variables such as team batting average, team slugging percentage, the potential bunter's slugging percentage and sacrifice bunting efficiency. It could also be applied to determine whether or not to pinch hit in such situations, comparing the batter's bunting to the potential pinch hitter's hitting or vice versa, or even to determine who to use to pinch hit, comparing one potential pinch hitter's bunting to another potential pinch hitter's hitting. It will not be concerned with trying to determine when the scoring of one run should be a primary objective since this is the decision of the manager. The output of the computer program will be used to generate two regression equations. This is a new approach to decision analysis since, to our knowledge, no one has attempted to do regression analysis on the output of a simulation of this type even though there would seem to be some significant advantages to doing this.

The first equation will equate the probability of scoring at least one run to a function of individual slugging percentage, team batting average and team slugging percentage, given that the player bats normally (P_1). The second equation will equate the probability of scoring at least one run to a function of individual sacrifice bunting efficiency, team batting average and team slugging percentage, given that the player attempts to sacrifice bunt (P_2). Both of these are conditional probabilities. They are conditional on the strategy used and on the values of team batting average and slugging percentage. Then, since the two equations are both probabilities of scoring at least one run, by subtracting the two equations, the resulting equation, when simplified, will relate sacrifice bunting efficiency to individual slugging percentage, team slugging percentage and team batting average. When given the values of the individual's slugging percentage, the team batting average and the team slugging percentage, this equation will generate the sacrifice bunting efficiency that would have an equal probability of scoring at least one run with that given by the other three values. This will allow us to compare the strategies by simply examining this equation. We determine the better strategy by computing the boundary value for sacrifice bunting efficiency by entering all the other values in the equation for a given situation. For a given individual with known bunting efficiency, we compare the boundary value with the known bunting efficieny. If the boundary value is less than the known value, it is better to sacrifice bunt. The two strategies can be compared graphically with the equation generated defining a curve in a plane, for fixed team batting average and team slugging percentage, where the probabilities involved are equal. Therefore, the better strategy for each player can be determined by considering the values of the variables for that player and observing which side of the curve these values indicate. This is similar to discriminant analysis in that the equation produced can be manipulated so that it is a type of discriminant function.

Discriminant analysis is the technique of using a number of measurements on some individual or object to classify that individual or object into one and only one of a number of categories. For our problem, we are trying to classify the situation into better to sacrifice bunt or better to bat normally based on the values of sacrifice bunting efficiency, player slugging percentage, team slugging percentage and team batting average. This can be done in this case by defining the discriminant function to be equal to the right side of the above equation minus sacrifice bunting efficiency. Therefore, if the discriminant function was less than zero, the situation would be classified as being better to sacrifice bunt which is consistent with the method above. Similarly, if the discriminant function was greater than zero, then the situation would be classified as being better to hit normally. The graphical analysis is also consistent with discriminant analysis. The difference is that the classification that is being done is based on simulated probabilities and not on some observable characteristic. It is not possible to determine into which category an individual should be classified by simply observing a single outcome since the fact that the strategy was unsuccessful does not necessarily mean that it was not still the better strategy. We would have to compare all the situations simulated and classify them ourselves based on the output, but, by doing a regression, this work is done for us. Also, in discriminant analysis, the discriminant function is linear, which was not the case here.

Using this method of performing regression analyses on the output of many simulations has a number of advantages. The simulation is random by design and, therefore, should automatically produce a random sample. Essentially, the simulation can be designed to satisfy most of the assumptions necessary to do a regression and produce meaningful results. If any statistical test indicates that an assumption has been violated, it is a problem of design and, in most cases, should be easily corrected, whereas, in most other investigations, the discovery that the sample is not random, for example, could indicate a massive waste of resources and a resolution by which randomness is attained may not be so easy to find.

It is also much easier to include the variables of interest, provided that these variables are sufficient to generate an adequate model, and exclude any other variables, while at the same time maintaining the desired general applicability. This is important when doing an investigation involving the sacrifice bunt because there are many variables which could have a significant effect on the probabilities involved. If one or more of these variables strongly suggest a course of action, then, naturally, these factors should be heeded. If, however, the situational variables do not suggest a course of action, it is then that these results become useful.

Previous methods of investigation have been of limited applicability for attempting to determine which strategy has the better probability of scoring at least one run. The first reason is that previous investigations having focused on the criteria of higher average number of runs scored instead of the probability of scoring at least one run. Other reasons for this are either the investigator has focused on the effect of using the sacrifice bunt on one specific lineup, which is only applicable to that lineup, or the investigator has used what has occurred throughout the recorded history of baseball to estimate the probabilities of interest. The investigator then applied various statistical or operations research methods to arrive at their conclusions. The first problem is that using one sacrifice bunting efficiency as every player's sacrifice bunting efficiency is inaccurate since not all players are equal in their ability to sacrifice bunt. Therefore, their results are only applicable to situations with players who have that same average ability to sacrifice bunt. Essentially, ability to sacrifice bunt has been ignored as a factor on the probability of scoring at least one run when the sacrifice bunt is used. The second problem is that to use what has occurred during the recorded history of baseball to estimate the desired probabilities, it would be desirable, first, to identify *all* the possible variables which could have a significant effect on the probabilities involved, second, to determine which of these variables actually have a significant effect on the probabilities of interest and whether that effect is consistent for all players, and, finally, consider only the cases for which these variables are neutral. The amount of work that would be involved would be overwhelming and we are not even sure it is possible to identify all the possible factors on the probability of getting a hit, let alone any other of the actions involved. The result of this would probably be that the number of cases which remained would not be large enough to produce confident estimates of the probabilities of interest, especially when you consider that with the sacrifice bunt, one can only count the cases where the ability to sacrifice bunt is comparable to the situation that is being considered. By using a simulation, we are able to concentrate on the variables of interest, provided that these variables are sufficient to produce an adequate model, and assume any other variables, significant or not significant, to be neutral. Therefore, we can easily generate a large enough sample to produce accurate estimates of the probabilities involved.

The method used is applicable to other situations involving decisions and determining the best course of action. The potential difficulties of using this method for other situations and the general characteristics necessary to use this method will also be considered. In chapter 6, we discuss these issues.

2 A Survey of O.R. and Statistical Analysis in Baseball

There have been a great number of articles written about mathematical and operations research applications to sports. Some useful references are listed in the survey article "General Review of OR in Sports" by Yigal Gerchak (1994). Articles which deal with applications to baseball are the most prevalent.

One of the most frequently explored topics has been strategy analysis in baseball. One article that considers this topic is "Analysis of Baseball as a Markov Process" by Richard Trueman (1977). In the article it was proposed that the game of baseball satisfied the four basic requirements of a Stationary Process:

- Baseball has a finite number of states.
- The transition probabilities do not change over time. (Trueman claimed that this is a slight simplification since it ignores situational and managerial strategies)
- The probability that a system will be in a state depends only on the previous state and not how that state was reached.

- For all states, the probability that the system initially occupies that state is known since there is only one possible initial state.

Using the everyday lineup of the 1973 Los Angeles Dodgers, Trueman showed how a detailed model of baseball, considered as a Markov Process, could be developed to evaluate different lineup orders and analyze strategies for many different situations. Trueman concluded:

- The attempted sacrifice was, in general, a very poor strategy.
- Only the pitcher should be asked to bunt, although, he admits that one or two others in the lineup could reasonably attempt to bunt, if they are excellent bunters.
- If the pitcher is a good bunter, he can try to sacrifice in the situation of a runner on first and one out.
- If the batter can successfully sacrifice 75 percent of the time, the suicide squeeze should be considered.
- Even in situations where the probability of scoring at least one run is increased, the sacrifice always reduces the expected number of runs scored.

- Basestealing can be worthwhile with good basestealers.
- The attempted steal is best if the baserunner is on first base. as compared to second or third base, with the strategy being better when there are more outs recorded as compared to less outs recorded.
- For a baserunner on third and two outs, stealing home plate is advisable only if successfully at least one third of the time.
- With less that two outs, the double steal will have a slightly lower required success probability.

Trueman also looked at lineup rearrangement to attempt to find the most productive lineup in terms of expected runs scored. Trueman actually found a lineup different from that commonly used with a slightly higher expected number of runs scored, but he did not believe that the improvement was significant enough to warrant a conclusion of rearranging the lineup.

Another article dealing with baseball strategy was "A Scientific Approach to Strategy in Baseball" by G. R. Lindsey (1977), in which Lindsey used probability theory and the statistics from the 1958, 1959 and 1960 Major League Baseball seasons to determine the probability of winning a baseball game given a lead or a deficit in a given inning, and to find the distribution of runs per half-inning. Using these results, Lindsey considered the strategies of the intentional walk, the sacrifice bunt and the stolen base to try to determine whether their use resulted in a significant difference in the probability of winning a game. For the sacrifice bunt and the stolen base, Lindsey's conclusions were essentially the same as Trueman's. For the intentional walk, Lindsey concluded that, in general, it was not a good strategy and the only time it might be useful was when there was only one out, a man on third base, the lead was one run and it is the ninth inning. This created the possibility of a game ending double play. It is acknowledged that, however, the intentional walk should be assessed on an individual basis. Lindsey also considered the value of outstanding players and a new 'batting efficiency' statistic which, by determining the increase or decrease of all possible results on the expected number of runs scored, could convert a player's results into a new statistic. Lindsey also noted the limitation of using past performances as a predictor of future performance.

In the article "Baseball a la Russe" by Ronald Howard (1977), which was also concerned with baseball strategies, Howard modeled baseball as a Markov process with 25 states and used Dynamic Programming to examine the effectiveness of various strategies. Howard considered strategies such as base stealing, the intentional walk and the sacrifice bunt as having a short term or immediate focus. He considered having a short term focus as being not as effective as having a long term focus, which basically meant letting each batter hit normally, when considering the expected number of runs scored. Howard calculated the probability of scoring x_i number of runs from state i using an average lineup and the expected number of runs scored in each situation, before and after the proposed strategy, to arrive at his conclusions.

Another article involving strategy analysis was "Dynamic Programming and Markovian Decision Processes, with Application to Baseball" by Richard Bellman (1977), which arrives at the same basic conclusions as the previous articles. The one difference is that Bellman acknowledges that the percentages are not a definitive standard upon which to base strategic decisions. As Bellman states:

> *As in poker, one can do very well playing the percentages; but if one wants to win big, one has to play psychology.*

The distribution of runs in a baseball game is another topic which has been considered. Two articles that looked at this topic are "The Distribution of Runs in the Game of Baseball" by D. A. D'Esopo and B. Lefkowitz (1977), and "Percentage Baseball, An Analysis of Baseball as a Game of Chance by the Monte Carlo Method" by Earnshaw Cook (1977), as well as Earnshaw Cook's book Percentage Baseball (1966). Both articles used probability theory to calculate the probabilities of scoring various numbers of runs in given situations. Both articles ignored situational variables. In Cook's book, he also looked at strategies as well as other topics, such as batting order. Cook's conclusions supported the obvious results, such as that the sacrifice bunt is not, *in general*, a useful strategy, but Cook believed his results to be novel and even went so far as to suggest that, by using his book and applying its principles, an average team would be elevated to becoming a pennant contender. Cook had difficulty understanding why no one involved with baseball was using his book. All other articles generally cautioned against applying the results too broadly, whereas, Cook seemed to openly invite this.

Batting order has been another topic considered in many articles, such as "Monte Carlo Analysis of Baseball Batting Order" by R. Allen Freeze (1977), and "Comparing the Run-Scoring Abilities of Two Different Batting Orders: Results of a Simulation" by Arthur Peterson (1977). Both articles used simulation to test different orderings of the same batters to observe which lineup would produce the most runs. Freeze used an existing New York Yankees lineup, as well as an all-star New York Yankees lineup as a basis, whereas Peterson used a fabricated average lineup. Both ignored situational variables and the fact that the batter's results would not necessarily have the same distribution for a given batter if that batter were hitting in a different position in the lineup. They both found some differences in the number of runs produced, but the new lineups were not significantly better that the standard lineups used.

Another topic that has been considered is playoff and tournament structures and elimination. One article that considered this topic is "On the Probability that the Better Team Wins the World Series" by James Kepner (1985). In this article, Kepner considered two questions. 'How many games must be played in an uncurtailed World Series so that we may be reasonably confident that the better team will win the World Series?' and 'How many games do we expect to be played in a curtailed World Series?'. For the first question, Kepner considered the variable $Y_i = 0$ or 1 depending on whether or not the better team wins the i^{th} game and then let $Y = \sum Yi$. Let $p = P(Y_i = 1)$ where the outcomes are independent. Kepner observed that Y is a binomially distributed random variable with mean np and variance

$np(1-p)$, where n is the number of games. Therefore, using the Central Limit Theorem, one can find $P(Y \geq n/2)$ for the desired confidence level. For the second question, Kepner found the expected value on N, the number of games necessary to win a curtailed World Series, where N could be any number between the minimum number of games necessary to win the World Series and n, the number of games in the World Series. Kepner then produced a distribution for $P(N = x)$ and found that the probability that the better team wins a curtailed World Series that may last as long as n games is $p^{y+1} \sum \frac{x!}{(y!)(x-y)!} q^{x-y}$ which is summed from $x = y$ to $x = n - 1$. Kepner evaluated this probability for different n and p to arrive at answers for different confidence levels. Kepner also found that the probability that the better team wins an uncurtailed World Series is equal to the probability that the better team wins a curtailed World Series.

Another article concerning playoff eliminations is "Baseball Playoff Elimination: An application of linear programming" by Lawrence Robinson (1991) which is a proposal for a better method for mathematically eliminating a team from playoff contention. The present system concludes that a team is eliminated when they are trailing the leading team or teams by more games than remain on the regular season schedule. This assumes that no other team can affect the outcome of whether any one team reaches the playoffs. Robinson proposed to use linear programming to determine for team A, the best possible outcomes for all remaining games not including team A, which might allow team A to reach the playoffs, assuming team A wins all of its remaining games. If team A cannot make the playoffs under this best possible scenario, then team A would be eliminated from playoff contention, and would do so at least as fast as the present system, and would probably be faster for most teams. For example, if team 1 is leading a division, team 2 is 2 games behind team 1 and team 3 is 8 games behind team 1, and there are 10 games remaining in the schedule, then under the present system, team 3 would not yet be eliminated from playoff contention since they trail team 1 by less games than remain in the schedule. Suppose, however, that team 1 and team 2 play each other in 7 of the last 10 games. The best result that team 3 could hope for is that team 2 wins 4 or 5 of these games which would put team 1 and team 2 within 1 game of each other at the end of the schedule. But this would mean that, at best, team 3 would have to win 11 games to tie the division leader, which is impossible since only 10 games remain. Therefore, team 3 would be eliminated under Robinson's system. Robinson's system is used on a limited basis, but he proposed a much broader application and even, possibly using the system in hockey and basketball.

In the article "Choice Models for Predicting Divisional Winners in Major League Baseball" by Daniel Barry and J. A. Hartigan (1993), a system was developed for calculating the probability that a team would win their division and, therefore, could predict a division winner based on the team with the highest probability of winning. The system estimated each team's probability of winning the division given what had occurred in the games played thus far, by producing a model for predicting which team would win each of the remaining games. The model depended on which teams are playing, which team is the home team and allows for different team strengths and home field advantages. Barry and Hartigan used Markov chain sampling to simulate the outcomes of future games, while at the same time, they made allowances for changing team strengths since teams will appear to change over the course of a season, as not all players have the same durability. This system was applied to the 1991 National League season at the all-star break and it correctly predicted that the Atlanta Braves would win the National League West division, even though the Braves trailed the Los Angeles Dodgers at that time. It would be interesting to see this model applied to other seasons to see how well it would perform.

Another article which deals with tournament structure is "Double-Elimination Tournaments: Counting and Calculating" by Christopher Edwards (1996). This article looked at different structures for single and double elimination tournaments and developed a method for determining the probability of winning a double elimination tournament. There are three possible scenarios for winning;

- Win the winner's tournament and then defeat the winner of the loser's tournament.
- Win the winner's tournament, lose once to the winner of the loser's tournament and then defeat the winner of the loser's tournament.
- Win the loser's tournament and then defeat the winner of the winner's tournament twice.

Given the probability of any one team defeating another, Edwards developed probabilities for winning in each round of a tournament, based on potential 'seating' in each round and then calculated the probability of winning a single elimination tournament. This could then be used to calculate the probability of winning a double elimination tournament. Edwards demonstrated this using a four team tournament. Edwards also looked at calculating the number of tournaments, structures and draws needed for a tournament involving a given number of teams playing and different types of structures.

Another important topic is player evaluation and its applications. One article on this topic is "Did Shoeless Joe Jackson Throw the 1919 World Series?" by Jay Bennett (1993). In this article, Bennett used the concepts of a Player Win Average and Player Game Percentage to show that, by the way Jackson played in the 1919 World Series, there is no reason to believe that he threw the World Series. The Player Win Average is based on the premise that the performance of baseball players should be quantified, based on the degree that their performance increased or decreased their team's chance of victory in each game. To find the Player Win Average (PWA), one must find $| \Delta WP |$ (the change in the probability of a win by the player's team) after a play involving the player, either offensive or defensive, and credit it to the player as Win Points if ΔWP is positive, or credit it to the player as Loss Points if ΔWP is negative, as originally proposed by Mills and Mills (1970). Then the PWA is equal to the total of the player's Win Points, divided by the total of the player's Win Points and Loss Points. The keys to this system is that it took the situation of a play into consideration, which is not done by normal baseball statistics, and that the PWA allowed for an effective way of comparing hitters and pitchers. PWA was especially good at evaluating relief pitchers and fielding performances in general. The major drawback of the PWA was that the calculation of the PWA was a greater data collection burden than standard baseball scorekeeping since it must determine the percentage of time each standard baseball event occurred and then thousands of games must be simulated to determine the win probabilities for each situation for all possible points in a game. Bennett used an updated Player Game Percent (PGP), which is based on the PWA, where $PGP = (WinPoints - LossPoints)/40$, (dividing by 40 provided a better scale). Bennett and Flueck also provided in their article "Player Game Percentage" (1984), a method for estimating win probabilities. One advantage of the PGP over the PWA is that it is easier to interpret since a positive PGP is good and a negative PGP is bad. The interpretation of the PWA is not so easy. The PGP also provides a more valid quantification of a player's contribution to victory. Another article on the topic of player evaluation is "The Valuation of a Baseball Player" by Carl Mitchell and Allen Michael (1977) which used simulation and Bayesian statistics to look at expenditure vs. expected return of having certain players on one's team.

A more in-depth examination of this topic from the point of view of salary was done in the article "Salary Evaluation for Professional Baseball Players" by James Lackritz (1990). The proposed evaluation was done using the effect of a player on the winning percentage of the team by measuring the impact of the performance statistics on the team's winning percentage and projecting this impact into dollars and cents and into home attendance. In previous unpublished papers, Lackritz proposed comparing a player with an 'average' player, with average being either a league or team average, and then multiplying any difference by the player's utilization function. The utilization function was defined as the player's total at-bats, fielding chances or innings pitched divided by his team's totals. This measured the player's fractional impact on his team's total chances. Lackritz established a base salary for players, based on how much they played, plus a bonus according to their final impact on the winning percentage, with it being possible to have a negative bonus. The statistics used were offensive average (equal to (total bases + walks + # of hit by pitch) / (total at-bats + walks)), on base percentage, stolen bases, ratio of strikeouts to walks allowed, hits per innings pitched, earned runs per innings pitched, # of saves per # of wins (this statistic compensated relief pitchers) and fielding percentage. For each of these statistics, one finds the difference between the player's statistic and the average, multiplied by the coefficient weight of the statistic calculated, multiplied by the appropriate utility function (offensive, defensive or pitching) to get a net effect on the percentage. Then one adds these numbers for all the statistics and multiples by 25000 dollars. This is the bonus to be added to the base salary. Lackritz also proposed a model for home attendance based on percentage of games won, last year's percentage of games won, the number of competing sports teams in the same market, weather effects on attendance, pennants won in the last 5 years and the number of 'superstars' on the team.

Another topic is the effect of the strike count on batting performance which was considered in the article "Batting Performance vs. Strike Count" by Pete Palmer (1977). Palmer used data from twelve World Series to calculate probabilities of success from different 'counts'. Palmer also looked at the effect of being ahead or behind in the count on batting performance. A related article was "A Statistical Analysis of Hitting Streaks in Baseball" by Christian Albright (1993), which looked at whether streaks occurred more often than expected under an assumption of randomness. This required the sequencing of success and failures instead of the totals of hits and at-bats. A previous study by the Elias Baseball Analyst found batting averages were just as likely to be higher following defined 'hot streaks' as following 'cold streaks'. Albright was concerned with how many players exhibited streaky sequences since some streakiness can be expected, and if hitters were perennially streaky or if streakiness was a one year phenomenon. The problem of the latter was that there were not enough streaky hitters to get significant results. The other problems were how to classify the results of successive at-bats (whether to distinguish the types of hits and how to classify walks), the effect of situational variables and how to define streakiness and randomness, in general. Albright used a method based on the number of 'runs,' a method of checking whether successive at-bats form a first-order Markov chain and a method using Logistic Regression Models. Albright failed to find convincing evidence of wide-scale streakiness.

3 Data and Simulation Methodology

The objective of this thesis is to develop a method to determine the strategy giving the higher probability of scoring at least one run- attempting a sacrifice bunt or batting normally (when the situational variables are essentially neutral and do not suggest a course of action). This will be done by simulating a large number of half innings in which each of the two options is used. There will be twelve different play results that will be considered by the simulation used in this thesis.

- No Advance: batter out, baserunners do not advance. (NoA)

- Sacrifice Out: batter out, baserunners all advance one base. (also deals with successful sacrifice bunts) (SO)

- Sacrifice Fly: batter out, baserunner on third base can score. (SF)

- Double Play: Play in which the batter and possibly one baserunner are forced out. If there is no baserunner on first base then there is one out recorded and no runners advance. If there is a baserunner on first base, then there are two outs recorded, except when there are baserunners on first and third only and no outs have been recorded. In that case, the baserunner on third is held at third, the baserunner on first advances to second base and the batter is out at first base. The reason is that getting the double play would have allowed the runner on third to score and so, holding the runner on third base means only one out can be recorded. Essentially, this is not really a double play. It is included in this category as a convenience. If the bases are loaded, then the baserunner at third is forced out at home plate and the batter is out at first base. (DP)

- Walk: batter to first base, baserunners advance if forced. (also includes hit batters) (Walk)

- ShortSingle: batter to first base, all baserunners advance one base only. (ShSgle)

- Single: batter to first base, baserunner on first base to second base and baserunners on second and third base score. (Sgle)

- Long Single: batter to first base, all baserunners advance two bases. (LgSgle)

- Short Double: batter to second base, all baserunners advance two bases.

(ShDble)

- Long Double: batter to second base, all baserunners score. (LgDble)
- Triple: batter to third base, all baserunners score. (Tple)
- Home Run: all baserunners and batter score. (HR)

Initially, the system has a baserunner on first base with no outs. The states of the system can be described as follows. For each base, there will be a player there or not. For three bases, this gives $2^3 = 8$ possibilities. Also, there can be zero, one, two, or three outs recorded or a run scored. If there are three outs or a run scored, the simulation of that half inning ends. Therefore, nobody is left on base, effectively, once either of these states are entered. Thus, there are $8 \times 3 + 2 = 26$ states. Therefore, the states are

- Three outs
- Run scored
- The following states exist for zero out, one out and two out (24 states)
  1. No one on base
  2. Baserunner on first base

3. Baserunner on second base

4. Baserunner on third base

5. Baserunners on first and second base

6. Baserunners on first and third base

7. Baserunners on second and third base

8. Baserunners on first, second and third base

Diagram2: Markov Transition Matrix

$$\begin{bmatrix} A_0 & A_1 & A_2 & A_3 \\ 0 & B_1 & B_2 & B_3 \\ 0 & 0 & C_2 & C_3 \\ 0 & 0 & 0 & D_3 \end{bmatrix}$$

where $A_0, A_1, A_2, B_1, B_2, C_2$ are 8×8, A_3, B_3, C_3 are 8×2, and D_3 is 2×2. The 0 entries are matrices of appropriate size. The subscript denotes the number of outs. Therefore, A_0 would denote the matrix containing the probabilities of beginning a play in some state with no outs and moving to another state with no outs. For our analysis, we assume $A_1 = A_2 = B_1 = B_2 = C_2$,

$A_3 = B_3 = C_3$. We have two versions of A_0 depending on which of the strategies we use. The probabilities which are the entries for these matrices depend on the team batting average and hit distribution.

Team batting average is the average for the rest of the batters in the lineup. It excludes the potential bunter, which is important since it maintains the independence of the potential bunter's statistics with respect to the team's statistics for the purposes of the simulation. Similarly, team slugging percentage is the slugging percentage for the rest of the batters in the lineup. Sacrifice bunting efficiency is exactly that, the probability that an attempted sacrifice bunt is successful. Batting average is the number of hits divided by the number of official at bats, where each type of hit is given equal weight. Slugging percentage is basically a weighted batting average where instead of giving each type of hit equal weight, each type of hit is given weight according to the number of bases that are reached. For example, a double is counted as two bases and a home run is counted as four bases.

The probability of each of these plays occurring will be based, initially, on their occurrence during the 1994 and 1995 Major League baseball seasons, which represents an average distribution of hits. This information was found in the STATS Player Profiles 1995 (1994) and the STATS Player Profiles 1996 (1995), as well as an existing baseball simulation game, Pursue the Pennant for the years 1995 and 1996. After running sufficient simulations for the initial distribution. simulations will then be run using probabilities generated from a distribution that results in a lower slugging percentage, and then a distribution that results in a higher slugging percentage from initial distribution.

The first batter, or potential bunter, will be assigned an average between 0.150 and 0.400. which is a reasonable range for a batting average of a player who one would not automatically replace with a pinch hitter, but would consider having sacrifice bunt. The slugging percentage is determined by the batting average and the distribution of hits. For the different batting averages and hit distributions used, the potential bunter's slugging percentage ranged between 0.177 and 0.86. The rest of the lineup will all have the same batting average, which will range between .248 and .308. This represents the majority of team batting averages. The rest of the lineup will have the same slugging percentage, which as a function of the batting average and the hit distribution, ranged between 0.29264 and 0.6192. Ideally, it would be best to perform the simulations with all batters having potentially different batting averages and slugging percentages. The problem with this approach is that it there are, effectively, an infinite number of combinations of batting averages and slugging percentages for comprising a single team batting average and team slugging percentage. The only other alternative would be to use a much more complex model that would have as many as twelve variables, and would still require an extraordinarily large number of simulations to generate a result and still not be confident that all possible combinations were included. Several thousand simulations were performed for the method that was used here.

The distribution of hits will be based, initially, on their distribution during the 1994 and 1995 seasons, which was found in the STATS Player Profiles 1995 and the STATS Player Profiles 1996. Therefore, if one out of every ten hits is a home run, then the batter will produce one home run for every 10 hits, regardless of batting average. The chance of a walk will be the same for every batter, which will be with the same frequency as it occurred during the 1994 and 1995 seasons. The chance of a double play, sacrifice out and sacrifice fly will be a function of the batter's chance of not getting a hit or walk and will also be based on the 1994 and 1995 seasons. For example, if one out of every ten outs would be a potential double play a double play would have been recorded if there were a runner on first base, then the double play result will occur ten percent of $(1 - (OnBasePercentage))$.

The simulation will be of a single half-inning and each simulated half-inning will start with a man on first and no outs. The simulation will observe how often at least one run is scored for the given conditions. If the sacrifice bunt is attempted, it can result in a successful sacrifice, an infield single, no runner advancement or a double play. A successful sacrifice is when the batter is out, but the runner on first base advances to second base. An infield single is described as a short single, just as the double play result and the no advancement result are as described above. The no advancement result could mean that the baserunner is out and the bunter is on first base, or that the bunter popped out or struck out. The probabilities of each of these occurring will be a function of the sacrifice bunting efficiency and not the assigned batting average of the bunter. These probabilities will also be based on the 1994 and 1995 Major League baseball seasons.

Since the simulation is testing how often at least one run is scored, the simulation will stop each test when either one run is scored or when there are three outs recorded, with precedence given to three outs being recorded. Therefore, it will never occur that a run will score on the same play as the third out is recorded. It is also unnecessary to consider whether the potential bunter will be hitting again, since this cannot occur before either the third out is recorded or at least one run is scored, thus ending the test.

The computer program which performed the simulation appears in Appendix A. It was written using the Pascal programming language. With this program, it is relatively simple to change the distribution of hits. Also, after an equation is established, simulation was used to test the equation for lineups that do not have all hitters, other than the potential bunter, hitting with the same batting average and distribution of hits. This will verify that the equation is applicable to essentially any lineup.

After the computer program generated several thousand simulated results, the output was run through a regression program. It was hoped that totally linear models would be adequate for both the bunting option and the batting normally option. Unfortunately, this was not the case. The linear model for the bunting option was found to be very adequate with an R^2 value of 0.965054. The T value for testing that the coefficient is not zero for the square of the sacrifice bunting efficiency term was 0.34. This is not significant and, therefore, for the bunting option, the linear model was adequate. For the batting normally option, the linear model had an R^2 value of 0.860883, which is fairly good. However, by including the square of the player slugging percentage term, the R^2 value of this model was 0.908341, which is a significant improvement. Also, the T value for testing that the coefficient of this squared term is not zero had a value of -48.99, which is significant and, therefore, the squared term contributes significantly to the model and should be included.

4 Assumptions

The first assumption to be dealt with is the normality of the distribution of the errors about the regression line. This was tested by generating histograms. The histograms of the simulations generated from these data are all clearly symmetric and appear bell shaped. Therefore, since the generated observations are symmetric, even if the data are not normally distributed, the robustness of the model makes the conclusions reasonable. Normality is assumed when testing for equality of variance. The histograms generated are shown in Appendix B.

The next set of assumptions concerns the error terms. It is assumed that the mean error terms associated with different effects are uncorrelated, and that the variance is constant and equal for different team batting averages and team slugging percentages. Both of these assumptions can be checked by the examination of the residual plots. The residual plots are in a band along the x-axis, which is the ideal shape, suggesting equal variance. Random simulation also guarantees that the mean error terms associated with the different effects are uncorrelated. Since team batting average and team slugging percentage are independent of sacrifice bunting efficiency and slugging percentage of the potential bunter. there should be no expectation that the mean error terms would be correlated. Also, Levene's test for equal variance was performed on a preliminary data set to determine if any transformations were required and it was found that for the batting option, the F-value from the data was 0.25 with degrees of freedom of 1, 2497; which is not significant. For the bunting option, the F-value from the data was 0.00 with degrees of freedom of 1, 2007; which is also not significant. Therefore, there is no evidence that the assumption of equal variance is violated. The residual plots generated are shown in Appendix C.

The assumption that the models used are adequate is tested. The results of these tests were summarized in the Methodology section. We used the R^2 value, which measures how much variation in the dependent variable can be accounted for by the model. A value close to 1 indicates a strong relationship and, therefore, a good fit for the model. The value of R^2 for the linear model for the bunting option was 0.965054, which indicated that the linear model provided a good fit for this model. The R^2 value for the model for the batting normally option was 0.908341, which also indicated a good fit for this model. As noted in the previous section, a linear model was hoped to be adequate for both models, and it was in the course of checking the adequacy of the linear model that a better model was discovered.

5 Results

The abbreviations used are

- P = Probability of Scoring at Least One Run (P_1 and P_2)
- TBA = Team Batting Average
- TSP = Team Slugging Percentage
- PSP = Player Slugging Percentage
- SBE = Sacrifice Bunting Efficiency

The equation generated for the batting option is

$$P_1 = -0.106279018 + (1.788996558 \times TBA) + (-0.306195021 \times TSP) + (0.6338697 \times PSP) - (0.359820521 \times PSP^2)$$

The equation generated for the bunting option is

$$P_2 = -0.096135634 + (1.226926907 \times TBA) + (0.032219952 \times TSP) + (0.251559797 \times SBE).$$

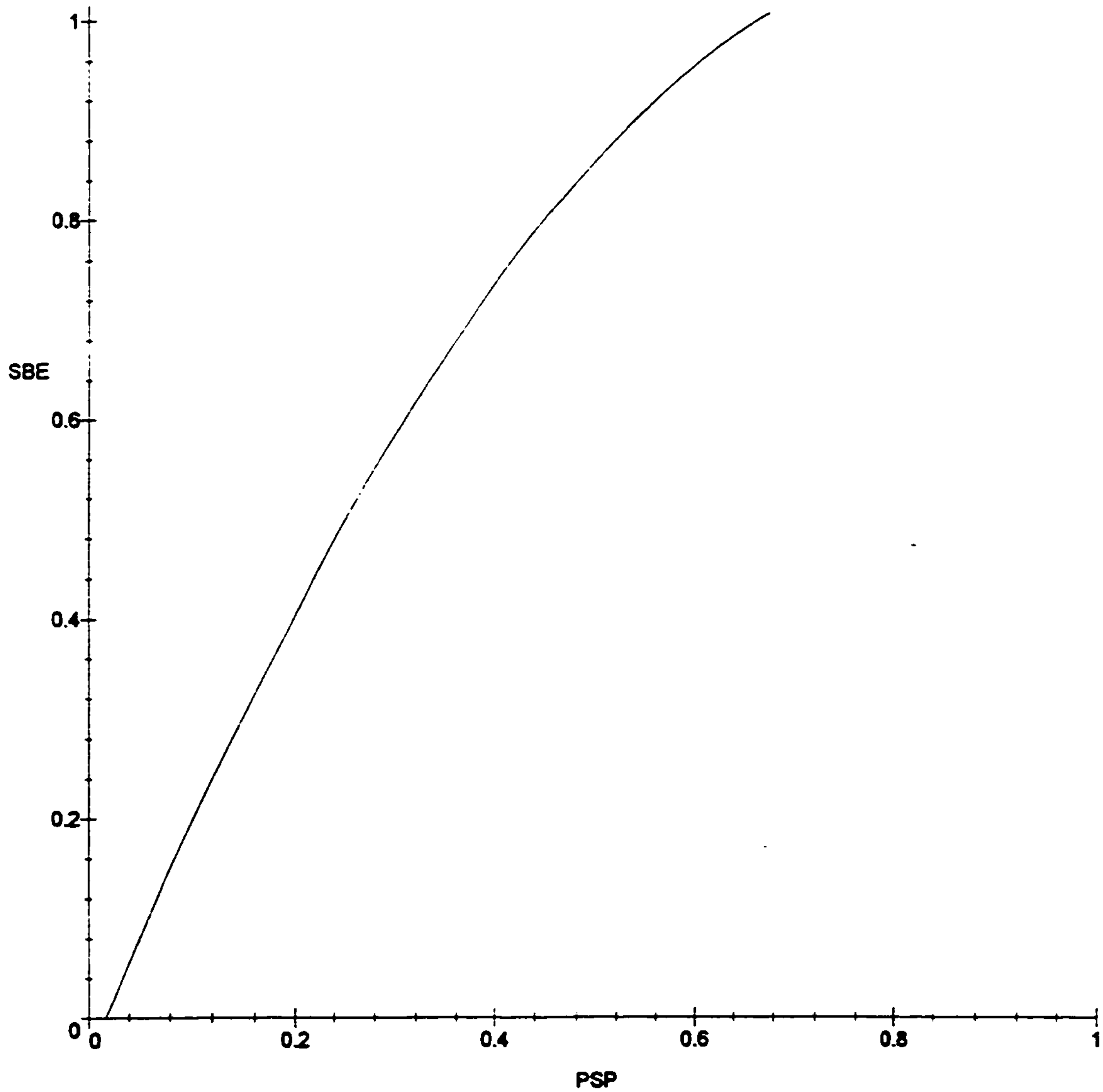
Therefore, when the two equations are subtracted and simplified, the resulting equation, which determines where P_1 and P_2 are equal, is

$$SBE = -0.04032 + (2.23434 \times TBA) - (1.34527 \times TSP) + (2.51976 \times PSP) -$$

($1.43036 \times PSP^2$). (1) Although the given equations could be less than zero for certain values, the range of values for the variables involved are such that this would not normally occur.

This is the graph of this function given TBA = .3 and TSP = .5.

Diagram3:



This equation will provide the mechanism of the decision process. For fixed TBA and TSP, the equation represents a quadratic in a plane with the axes being SBE and PSP. This curve marks a boundary where P is equal for the two options. It is better to sacrifice bunt on one side, and on the other side, the situation is classified as better to bat normally. By entering the values of the potential bunter's slugging percentage, the team slugging percentage and the team batting average for the given situation when a sacrifice bunt is being considered, the equation will generate the sacrifice bunting efficiency which would have equal probability of scoring at least one run for the given values. Therefore, if the actual sacrifice bunting efficiency is greater than the generated value, the better strategy would be to sacrifice bunt. If the generated value is greater than the actual value of the player's sacrifice bunting efficiency, then that would indicate that letting the batter hit normally has a greater probability of scoring at least one run. For example, if equation (1) generated a sacrifice bunting efficency of .85 for a given situation, but the actual sacrifice bunting efiiciency was only .8, this would indicate that batting normally would be the better strategy for this situation.

Equation (1) was then tested for lineups that did not have all batters assigned the same batting average or hit distribution. After doing a large number of tests using this equation and comparing the results to those generated by simulation, it was found that the equation deals with non-homogeneous lineups fairly well. Here are ten examples from the many tests done.

1. $TBA = 0.27625$

   $TSP = 0.4455625$

   $PSP = .354$

   $SBE = .99$

   The simulation generated probabilities of 0.4186 for bunting and 0.3726 for batting normally. Therefore, simulation indicates that bunting is the better option. By entering the given values into equation (1), the actual SBE is greater than the generated value for SBE. Therefore, equation (1) also indicates that bunting is the better option, which agrees with the simulation result.

2. $TBA = 0.245$

   $TSP = 0.3897625$

   $PSP = 0.5375$

$SBE = 0.9$

The simulation generated probabilities of 0.4084 for bunting and 0.4142 for batting normally. Therefore, simulation indicates that batting normally is the better option. By entering the values into equation (1), the actual SBE is less than the generated value for SBE. Therefore, equation (1) also indicates that batting normally is the better option, which agrees with the simulation result.

3. $TBA = 0.23$

   $TSP = 0.3182$

   $PSP = 0.43175$

   $SBE = 0.3$

   The simulation generated probabilities of 0.2904 for bunting and 0.3796 for batting normally. Therefore, simulation indicates that batting normally is the better option. By entering the values into equation (1), the actual SBE is less than the generated value for SBE. Therefore, equation (1) also indicates that batting normally is the better option, which agrees with the simulation result.

4. $TBA = 0.27$

$TSP = 0.475875$

$PSP = 0.46315$

$SBE = 0.75$

The simulation generated probabilities of 0.4092 for bunting and 0.4178 for batting normally. Therefore, simulation indicates that batting normally is the better option. By entering the values into equation (1), the actual SBE is less than the generated value for SBE. Therefore, equation (1) also indicates that batting normally is the better option, which agrees with the simulation result.

5. $TBA = 0.25$

   $TSP = 0.378125$

   $PSP = 0.54825$

   $SBE = 0.95$

   The simulation generated probabilities of 0.412 for bunting and 0.4026 for batting normally. Therefore, simulation indicates that bunting is the better option. By entering the values into equation (1), the actual SBE is less than the generated value for SBE. Therefore, equation (1) indicates that batting normally is the better option, which disagrees with the simulation result. This indicates that equation (1) is not perfect or that the simulation is giving a false result. (The actual SBE is actually very close in value to the generated SBE)

6. $TBA = 0.275$

   $TSP = 0.375175$

   $PSP = 0.3032$

   $SBE = 0.7$

   The simulation generated probabilities of 0.3806 for bunting and 0.3996 for batting normally. Therefore, simulation indicates that batting normally is the better option. By entering the values into equation (1), the actual SBE is greater than the generated value for SBE. Therefore, equation (1) indicates that bunting is the better option, which disagrees with the simulated result. This indicates that equation (1) is not perfect or that the simulation is giving a false result. (The actual SBE is actually very close in value to the generated SBE)

7. $TBA = 0.2375$

   $TSP = 0.38015$

   $PSP = 0.46786$

$SBE = 0.8$

The simulation generated probabilities of 0.3792 for bunting and 0.3952 for batting normally. Therefore, simulation indicates that batting normally is the better option. By entering the values into equation (1), the actual SBE is less than the generated value for SBE. Therefore, equation (1) also indicates that batting normally is the better option, which agrees with the simulated result.

8. $TBA = 0.25$

   $TSP = 0.4410575$

   $PSP = 0.354$

   $SBE = 0.9$

   The simulation generated probabilities of 0.402 for bunting and 0.3822 for batting normally. Therefore, simulation indicates that bunting is the better option. By entering the values into equation (1), the actual SBE is greater than the generated value for SBE. Therefore, equation (1) also indicates that bunting is the better option, which agrees with the simulated result.

9. $TBA = 0.275$

$TSP = 0.4046875$

$PSP = 0.43$

$SBE = 0.7$

The simulation generated probabilities of 0.4012 for bunting and 0.4362 for batting normally. Therefore, simulation indicates that batting normally is the better option. By entering the values into equation (1), the actual SBE is less than the generated value for SBE. Therefore, equation (1) also indicates that battting normally is the better option.

10. $TBA = 0.2795$

    $TSP = 0.416035$

    $PSP = 0.471$

    $SBE = 0.97$

    The simulation generated probabilities of 0.4484 for bunting and 0.4226 for batting normally. Therefore, simulation indicates that bunting is the better option. By entering the values into equation (1), the actual SBE is greater than the generated value for SBE. Therefore, equation (1) also indicates that bunting is the better option.

These examples are fairly representative of all the tests done. The equation developed performs well for these lineups, although there are a small number of cases where the difference is very minor.

The next step is to consider whether or not to employ a minimum difference between the two sides of the equations when determining whether to sacrifice bunt. This means that if the difference in the values of the actual sacrifice bunting efficiency and the generated sacrifice bunting efficiency is greater than the minimum difference, then you would proceed as indicated. If the difference in values is less than the minimum difference, then it makes almost no difference probabilistically as to which strategy is used. Therefore, when there is no advantage to either strategy with respect to trying to score at least one run, a secondary consideration is which strategy scores a higher average number of runs. It has been shown in many articles that batting normally will score a higher average of runs. Since this is the case, the appropriate course of action would be to let the batter hit normally if the generated sacrifice bunting efficiency is greater than actual sacrifice bunting efficiency or if the actual sacrifice bunting efficiency is greater than the generated sacrifice bunting efficiency by a value which is less than the minimum difference. Therefore, the equation can be modified to reflect this simply by adding the appropriate amount to the right side of the equation equal to the value of the minimum difference. The value of the minimum difference can be subjective, decided upon by the manager in question, which can be zero, or it can be the value used for determining a confidence interval of the probability of scoring at least one run from the swinging away option. Essentially, this is a hypothesis test that the probability of scoring at least one run from the bunting option is different than the probability of scoring at least one run from the batting normally. From the output, the standard error of prediction for the batting option is estimated to be 0.01415. Using the value 2.326 for the value of $t_{0.99,4638}$, a confidence interval would be the generated predicted value $\pm$ 0.033 approximately.

Using this adjustment, all ten test cases above would produce results using the equation that are consistent with the simulated result. This does not mean that the equation perfectly produces the best decision. There are still cases for which the simulated probabilities disagree with the prediction by the equation, but the performance is improved. Different values for t can be used corresponding to different confidence intervals.

6 Other Applications

There are a number of other possible applications for the method we developed for determining the better baseball strategy. Essentially, this method provides a way to decide which is the best initial strategy in a given situation, given certain variables. The basic characteristics that are needed in any such situation where the goal is to find the strategy with the better probability are that:

- There can be only a finite number of attempts to achieve the desired objective. This could be a time limit as well as simply a limit on the number of attempts.

- There should be a standard procedure or strategy that is applied, as well as some manner of 'safety play,' or alternative strategy.

- The aim of the safety play is to make it easier to achieve the desired objective and will not, by itself, immediately achieve the desired objective. If the safety play is successful, then it will be easier to achieve the objective, and if not successful could make it harder.

- All the possible outcomes of a course of action must be known.

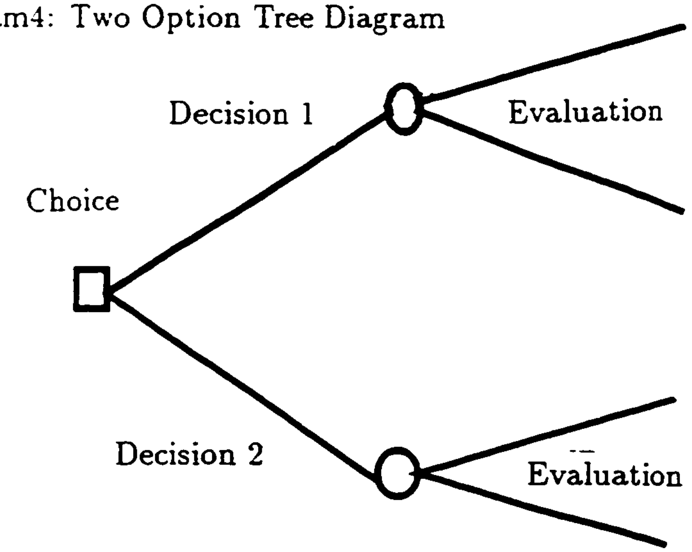
- There must be a method to obtain reasonable estimates of the probabilities involved.

- The variables of interest must be sufficient to provide an adequate model.

If the goal is not based on probability, but instead on cost or quickness, then the first, second and third items above do not apply. The fourth, fifth and sixth items are still relevant, and if cost or time is involved, there must be a method to obtain reasonable estimates of these quantities as well.

In our case, there are only two options that are being considered. This is the easiest situation since it guarantees that there will be only one equation generated for the purpose of comparison.

Diagram4: Two Option Tree Diagram



The method works well for this analysis because the model for the bunting option is linear and, therefore, the resulting equation is only as complex as the model for the batting normally option. This will not always be the case, although, ideally, both models will be linear and, therefore, will be a single line. We now consider a collection of general problems of a similar type. Let z be some measure of the success of some strategy1. Let x be a variable which influences z. Assume $z = f(x)$ in an equation relating the two variables. Similarly, we have $z = g(y)$ where y is a different variable and z is the same measure applied to some strategy2. By eliminating z from the two equations $z = f(x)$ and $z = g(y)$, we obtain an equation of the form $y = h(x)$ (this implicitly assumes that $g(y)$ is invertable, which is not always the case). This represents a boundary for which the two strategies are equivalent. It is interesting to consider what the boundary equation looks like for different functions $f(x)$ and $g(y)$. If the two regression lines are non-linear, non-monotone lines, the resulting equation can produce a very complex situation. It is possible to have many boundary lines and alternating areas of classification.

<u>Case1</u>

$z = g(y) = y - y^2$ (See Diagram5)

$z = f(x) = .5(\sin(x) + 1)$ (see Diagram6)

$y = h(x) = .5 - \sqrt{-.5\sin(x) - .25}$ (See Diagram7)

Diagram5: $z = g(y) = y - y^2$

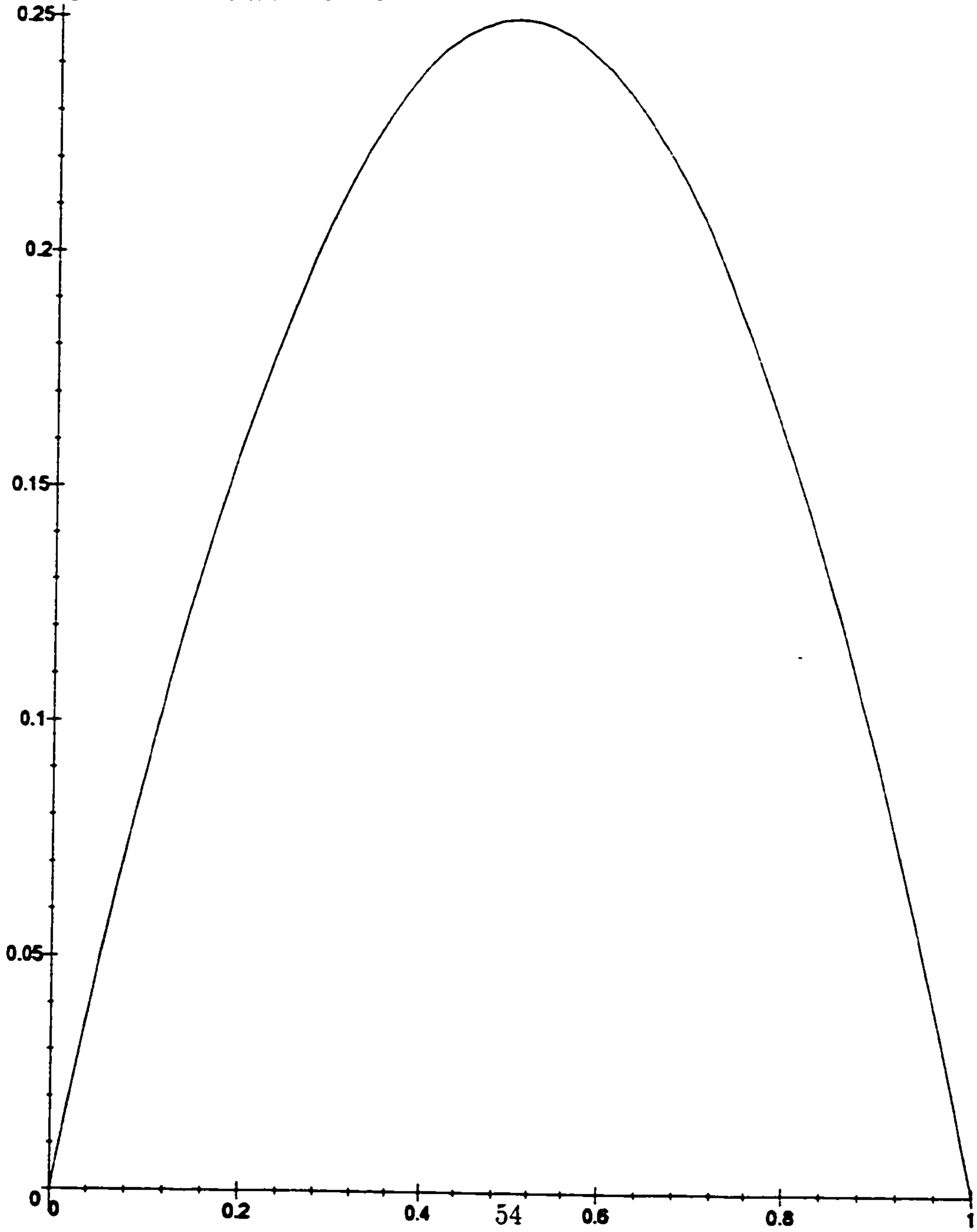


Diagram6: $z = f(x) = .5(\sin(x) + 1)$

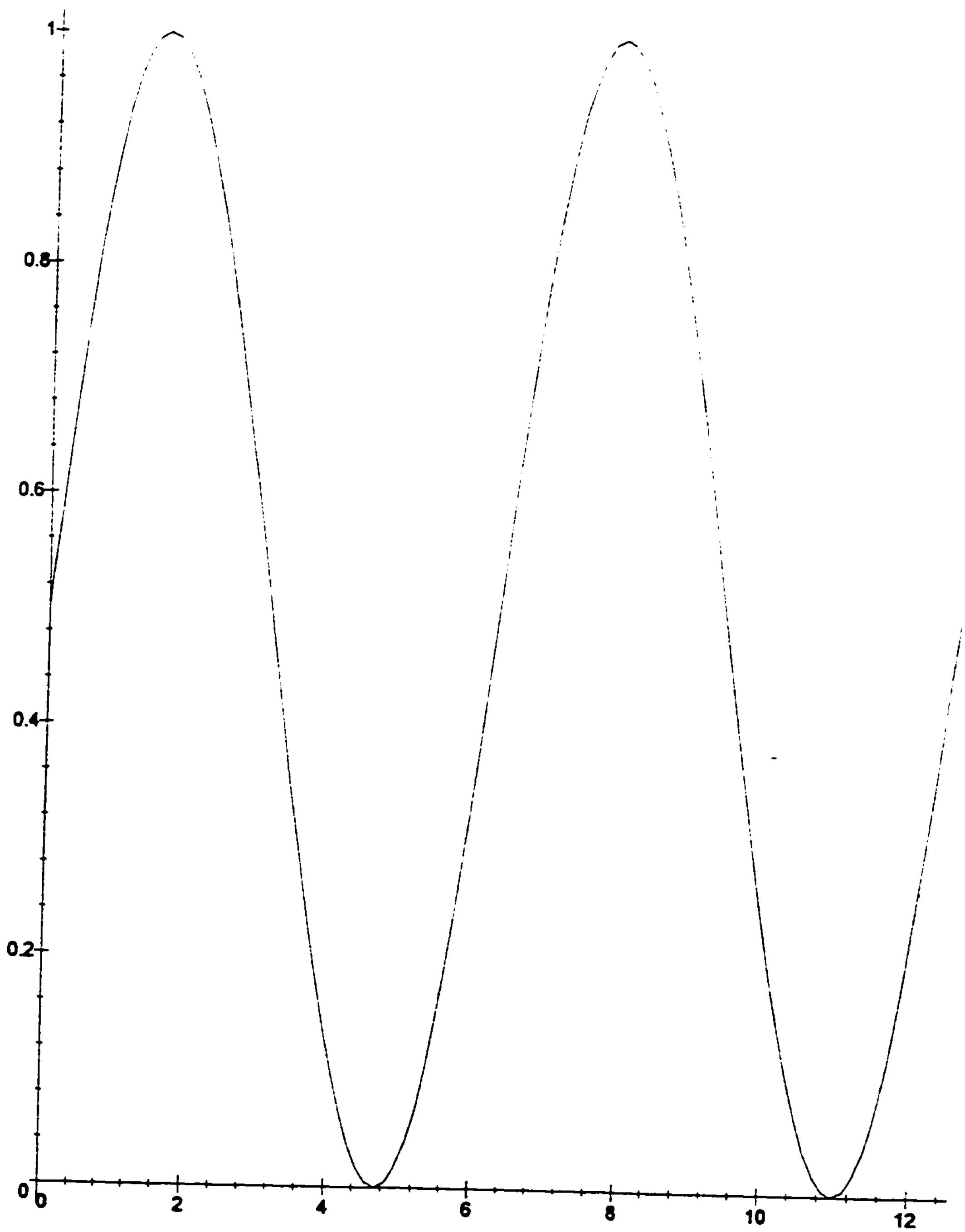
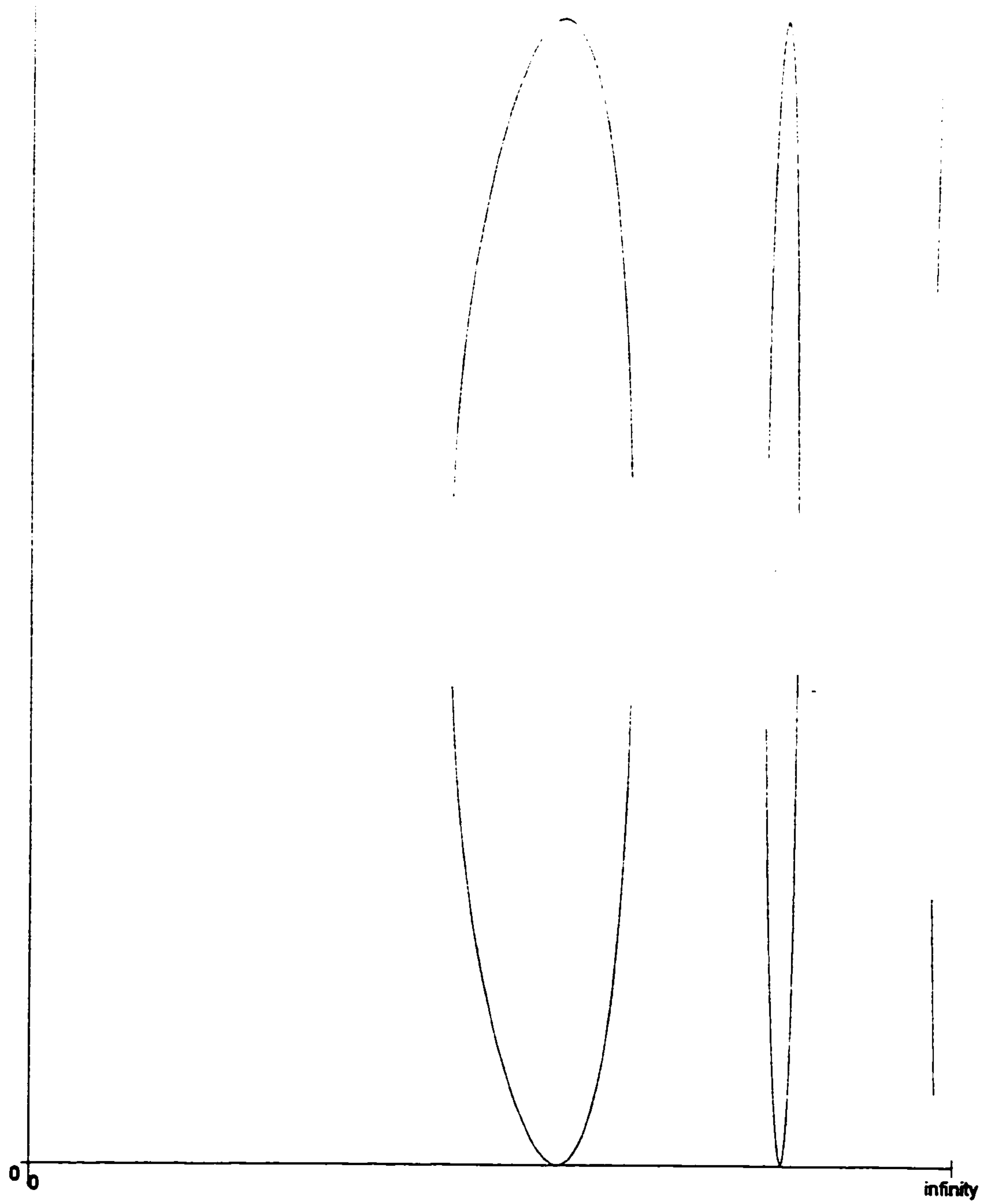


Diagram7: $y = h(x) = .5 - \sqrt{-.5\sin(x) - .25}$

The result is very complex.



Case2

$z = f(x) = x - x^2$ (See Diagram8)

$z = g(y) = y^2 - y + .25$ (See Diagram9)

$y = h(x) = .5 + \sqrt{x - x^2}$ (See Diagram10)

Diagram8: $z = f(x) = x - x^2$

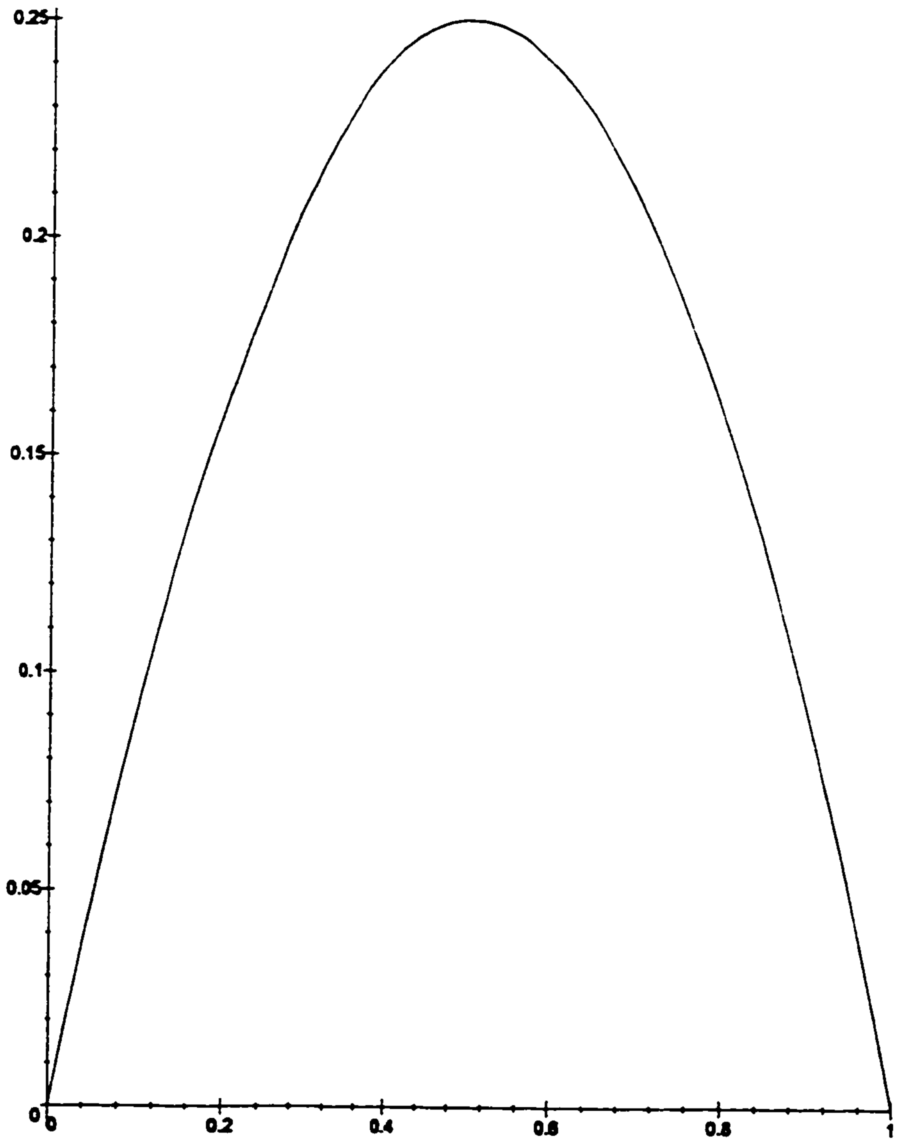


Diagram10: $z = g(y) = y^2 - y + .25$

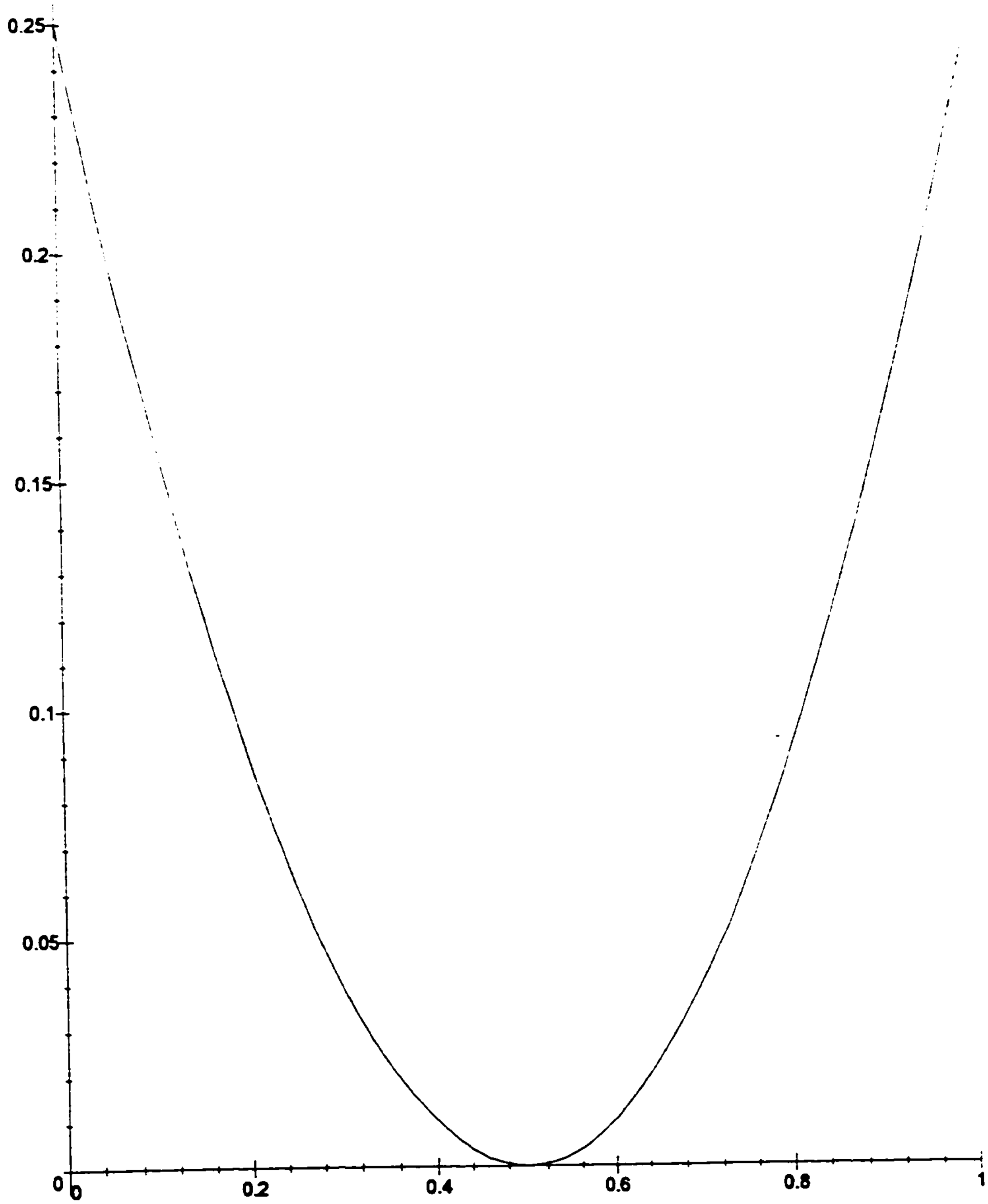
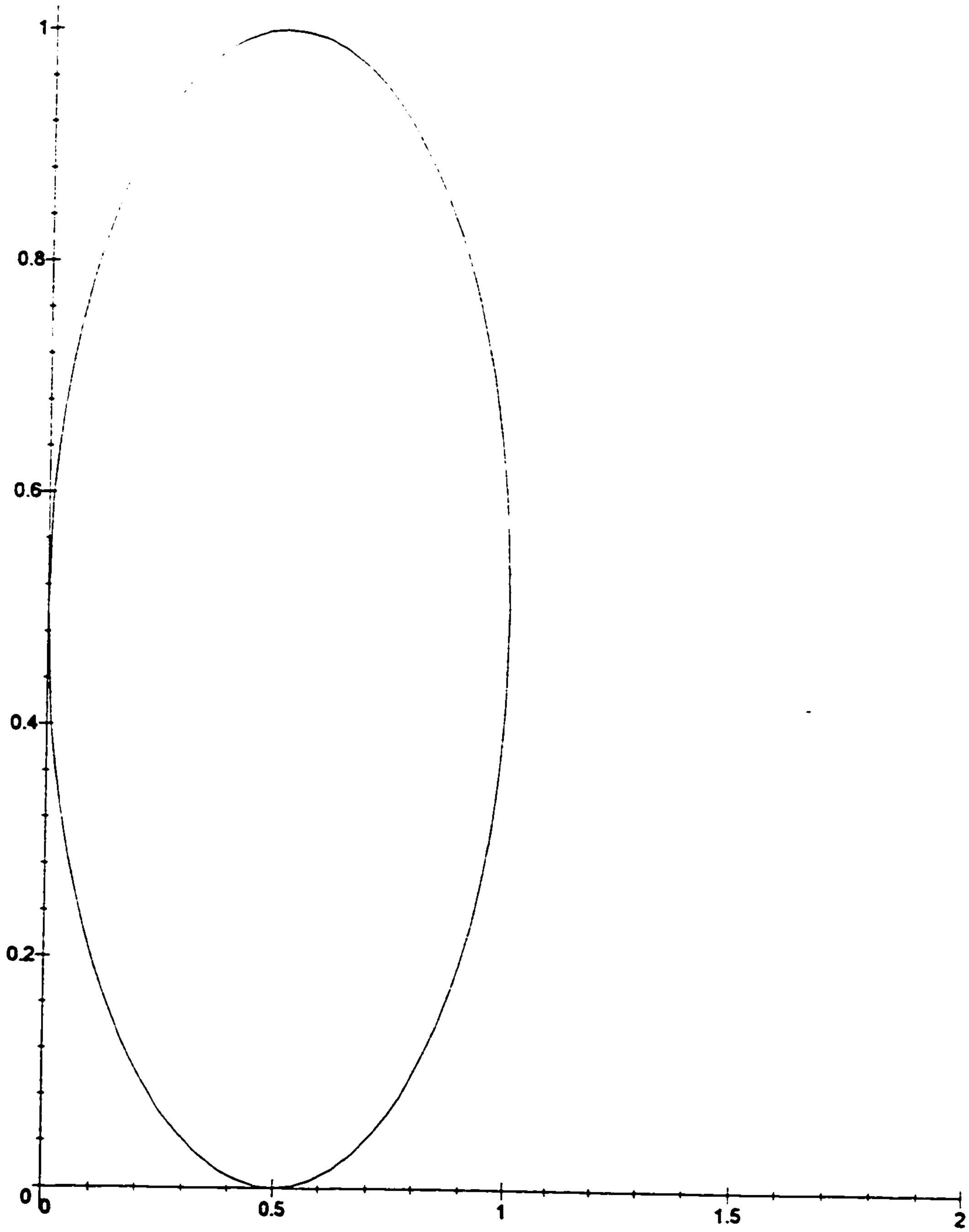


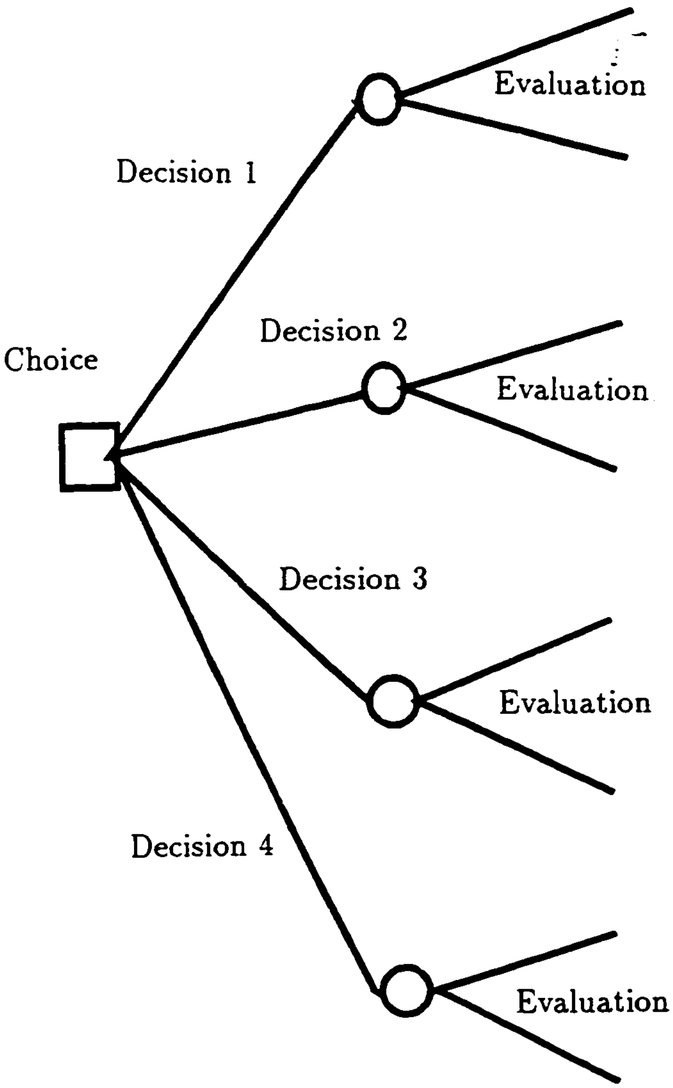
Diagram10: $y = h(x) = .5 + \sqrt{x - x^2}$

This is still relatively simple.



This method can also deal with situations for which the decision to be made has more than two options, but it becomes much more complex since, if there are n options, then this method can produce $\sum(n-1)$ equations for the purpose of comparison.

Diagram11: Multiple Options Tree Diagram



There are other applications in sports. In baseball, this method could be used for comparing bunting to batting normally when considering the average number of runs scored. This method could also be used for testing the strategies in baseball of base stealing, the hit-and-run play or the intentional walk. Of these three, the hit-and-run play would be the hardest to test since it involves probabilities not easily estimated, such as the probability of making contact on a swing.

Our method should be applicable in football, for deciding whether to punt or try a field goal given your field position and the score of the game, or to find the point at which it becomes more advantageous to try to score a touchdown instead of simply seeking to achieve a first down. For the field goal situation, the finite opportunities would be the time limit and the desired objective would be to win the game. The field goal, if successful, could give the team a lead such that the opposing team must either score twice (a lead of nine points or more) or must score a touchdown instead of a field goal (a lead of four to six points). If the field goal was not successful, this would give the opposing team good field position. By punting, the field position of the opposing team could be made much worse, hopefully within ten yards of their own end zone, thus making it much harder for the opposing team to score. The variables of interest are the field position, the score, 'field goal' ability, 'punting' ability and the relative offensive and defensive skills of the opposing teams.

It has been suggested that this method may be applicable in certain track and field events, such as high jump. This method may not apply to high jump, because of the problem of having only finite chances. In high jump, human endurance would seem to dictate that there would be only a finite number of chances to achieve the objective, but it would be hard to determine how many chances would be needed during the course of an event, especially since it could change from event to event.

Another possible application of our method is to find the strategy with the lower cost (instead of higher probability) in the following queueing problem. Assume a GI/G/2 queue, for specified GI and G. Suppose two "paired" arrivals (such as husband and wife) enter this queueing situation with two servers and equal two noempty lines. The "paired" arrivals have two possible strategies. One strategy is that they can each enter a line. When the first person enters service, the other person would renege. The other strategy is to have the "paired" arrivals stay together by choosing a line and both entering it. The service rates are known to be μ_1 and μ_2, where μ_1 is the higher rate. There are values attached to staying together, a per unit time. and values attached to being apart, b per unit time. We assume that the value of a is less than the value of b. This assumes that the pair prefers to be together. Splitting up is faster, on average, but has a higher cost per unit time. By doing a simulation and performing regression analyses on the output, the result should be two equations. The first equation, associated with the strategy of staying together, would equate cost to a function of a and μ_1. The second equation, associated with the strategy of splitting up, would equate cost to a function of b, μ_1 and μ_2. By combining the two equations and simplifying, the result would be an equation where a is equated to a function of b, μ_1 and μ_2 (could also compare by costs a and b). The values of μ_1, μ_2, a, b could be known from previous experience. The outcomes are all known and the variables should be sufficient for this model.

There are other possible applications in the world of medicine. The important point is to consider is that the baseball situation examined could be looked at as having two absorbing states (run scored and three outs) and all other situations are transient states collasped to the number of outs (none out, one out, and two out). When considered this way, the desired objective can be looked upon as simply trying to remain outside of the absorbing state of three outs. Similarly, there are many situations in medicine where one is simply trying to remain outside of some absorbing state, such as some stage of an illness which is considered to be irreversible.

One possible application is in the treatment of patients whose heart has stopped. The standard procedure is the application of electrical shocks or a defibrillator (which itself is a perfect example of a situational strategy), with a possible 'safety play' being the injection of adrenaline. The objective is to determine whether it is always best to administer adrenaline. The finite limit would be the time limit of getting the heart restarted before brain damage occurs. The situation variables could be the general health of the patient, the skill of the people providing the treatment and the quality of the equipment being used to provide the treatment.

Another possible application is determining the best initial treatment for people diagnosed with cancer. The options would be chemotherapy, as opposed to surgery or other possible treatments. The finite limit is the apparent time span after which the life expectancy of a person diagnosed with cancer does not differ significantly from a 'normal' person, and so, if the patient survives for this apparent time span after being diagnosed, the patient is considered to be cured. Again, the situational variables could be the general health of the patient, the skill of the people who would be providing the different treatments and the quality of the facility.

Another possible application is in the treatment of HIV positive patients. A patient does not want to enter the absorbing state of having AIDS. There is no real finite limit on the objective. There are a variety of treatments that are presently being used. This is another possible application in the field of medicine. Many others possibly exist.

7 Conclusion

The method of performing a regression on the output of a simulation is, potentially, a very useful method. The method of equating a common variable, measuring two possible actions, gives a useful boundary to describe the best choice. As applied to baseball, we have shown that, although the sacrifice bunt is not a generally useful strategy, it can be a useful strategy in the proper circumstances.

8 Appendix A

Program Baseball.pas

```
Program Baseball(output);

Var outs, basel, base2, base3, runs, count : Integer;

    hit, bunt, average, averageb : Real;


procedure Walk;

begin

if basel = 1 then begin

    if (base2 = 1) and (base3 = 1) then base3 := 2

        else begin

            if base2 = 1 then base3 := 1

            base2 := 1;

        end {else};

    end;

basel := 1;

end {Walk};
```

```
procedure ShSgle;
begin
if base3 = 1 then base3 := 2
    else begin
        if (base2 = 1) and (base1 = 0) then begin
            base3 := 1;
            base2 := 0;
        end
        else Walk
    end {else};
base1 := 1;
end {ShSgle};


procedure Sgle;
begin
if (base3 = 1) or (base2 = 1) then base3 := 2
    else begin
        if base1 then base2 := 1
```

```
    end {else};
base1 := 1;
end {Sgle};


procedure LgSgle;
begin
if (base3 = 1) or (base2 = 1) then base3 := 2
    else begin
        if base1 = 1 then base3 := 1
    end {else};
base1 := 1;
end {LgSgle};


procedure ShDble;
begin
LgSgle;
base1 := 0;
base2 := 1;
end {ShDble};
```

```
procedure LgDble;
begin
if base1 = 0 then ShDble
    else base3 := 2;
end {LgDble};

procedure Tple;
begin
LgDble;
if base3 = 0 then base3 := 1;
end {Tple};

procedure HR;
begin
base3 := 2;
end {HR};

procedure SF;
```

```
begin
if (outs < 2) and (base3 = 1) then base3 := 2;
outs := outs + 1;
end {SF};

procedure SO:
begin
if outs < 2 then ShSgle;
base1 := 0;
outs := outs + 1;
end {SO};

procedure NoA;
begin
outs := outs + 1;
end {NoA}

procedure DP;
begin
```

```
if base1 = 0 then NoA
    else begin
        if outs > 0 then outs := 3
            else begin
                if base3 = 1 then begin
                    if base2 = 1 then outs := 2
                        else begin
                            outs := 1
                            base2 := 1
                        end {else}
                    end
                    else begin
                        SO;
                        base2 := 0;
                        outs := 2;
                    end {else};
            end {else};
        end {else};
base1 := 0;
```

```
end {DP};

begin {main}
writeln('Enter team average in decimal form');
readln(average);
writeln('Enter bunter batting average in decimal form');
readln(averageb);
writeln('Enter sacrifice bunt efficiency in decimal form');
writeln('or zero if not bunting');
readln(bunt);
runs := 0;
Randomize;
for count : 1 to 5000
    do begin
        base1 := 1;
        base2 := 0;
        base3 := 0;
        outs := 0;
        if bunt > 0 then begin
```

```
hit := Random;
if hit < (.15 × bunt) then ShSgle
    else begin
        if hit < bunt then SO
            else begin
                if hit < (.2 × bunt + .8) then NoA
                    else DP;
            end {else};
    end {else};
end;
else begin
    hit := Random;
    if hit ≤ (.11 × averageb) then HR;
    if hit ≤ (.135 × averageb) and hit > (.11 × averageb)
    then Tple;
    if hit ≤ (.23 × averageb) and hit > (.135 × averageb)
    then LgDble;
    if hit ≤ (.325 × averageb) and hit > (.23 × averageb)
    then ShDble;
```

if hit $\leq$ (.55 $\times$ *averageb*) and hit $>$ (.325 $\times$ *averageb*)

then LgSgle;

if hit $\leq$ (.925 $\times$ *averageb*) and hit $>$ (.55 $\times$ *averageb*) then Sgle;

if hit $\leq$ *averageb* and hit $>$ (.925 $\times$ *averageb*) then ShSgle;

if hit $\leq$ (*averageb* + .06) and hit $>$ *averageb* then Walk;

if hit $\leq$ (.85 $\times$ *averageb* + .201) and hit $>$ (*averageb* + .06)

then SO;

if hit $\leq$ (.6 $\times$ *averageb* + .436) and hit $>$ (.85 $\times$ *averageb* + .201)

then SF;

if hit $\leq$ (.1 $\times$ *averageb* + .906) and hit $>$ (.6 $\times$ *averageb* + .436)

then NoA;

if hit $>$ (.1 $\times$ *averageb* + .906) then DP;

end {else};

repeat

hit := Random;

if hit $\leq$ (.11 $\times$ *average*) then HR;

if hit $\leq$ (.135 $\times$ *average*) and hit $>$ (.11 $\times$ *average*) then Tple;

if hit $\leq$ (.23 $\times$ *average*) and hit $>$ (.135 $\times$ *average*) then LgDble;

if hit $\leq$ (.325 $\times$ *average*) and hit $>$ (.23 $\times$ *average*) then ShDble;

if hit $\leq (.55 \times \mathit{average})$ and hit $> (.325 \times \mathit{average})$ then LgSgle;

if hit $\leq (.925 \times \mathit{average})$ and hit $> (.55 \times \mathit{average})$ then Sgle;

if hit $\leq \mathit{average}$ and hit $> (.925 \times \mathit{average})$ then ShSgle;

if hit $\leq (\mathit{average} + .06)$ and hit $> \mathit{average}$ then Walk;

if hit $\leq (.85 \times \mathit{average} + .201)$ and hit $> (\mathit{average} + .06)$ then SO;

if hit $\leq (.6 \times \mathit{average} + .436)$ and hit $> (.85 \times \mathit{average} + .201)$

then SF;

if hit $\leq (.1 \times \mathit{average} + .906)$ and hit $> (.6 \times \mathit{average} + .436)$

then NoA;

if hit $> (.1 \times \mathit{average} + .906)$ then DP;

until (base3 = 2) or (outs = 3);

if base3 = 2 then runs := runs + 1;

end {for};

if bunt = 0 then writeln('with no bunt, the # of runs was', runs, averageb);

else writeln('with the bunt, the # of runs was', runs, bunt);

end {main}

9 Appendix B

Diagram12:

This is the histogram generated from the batting normally option.

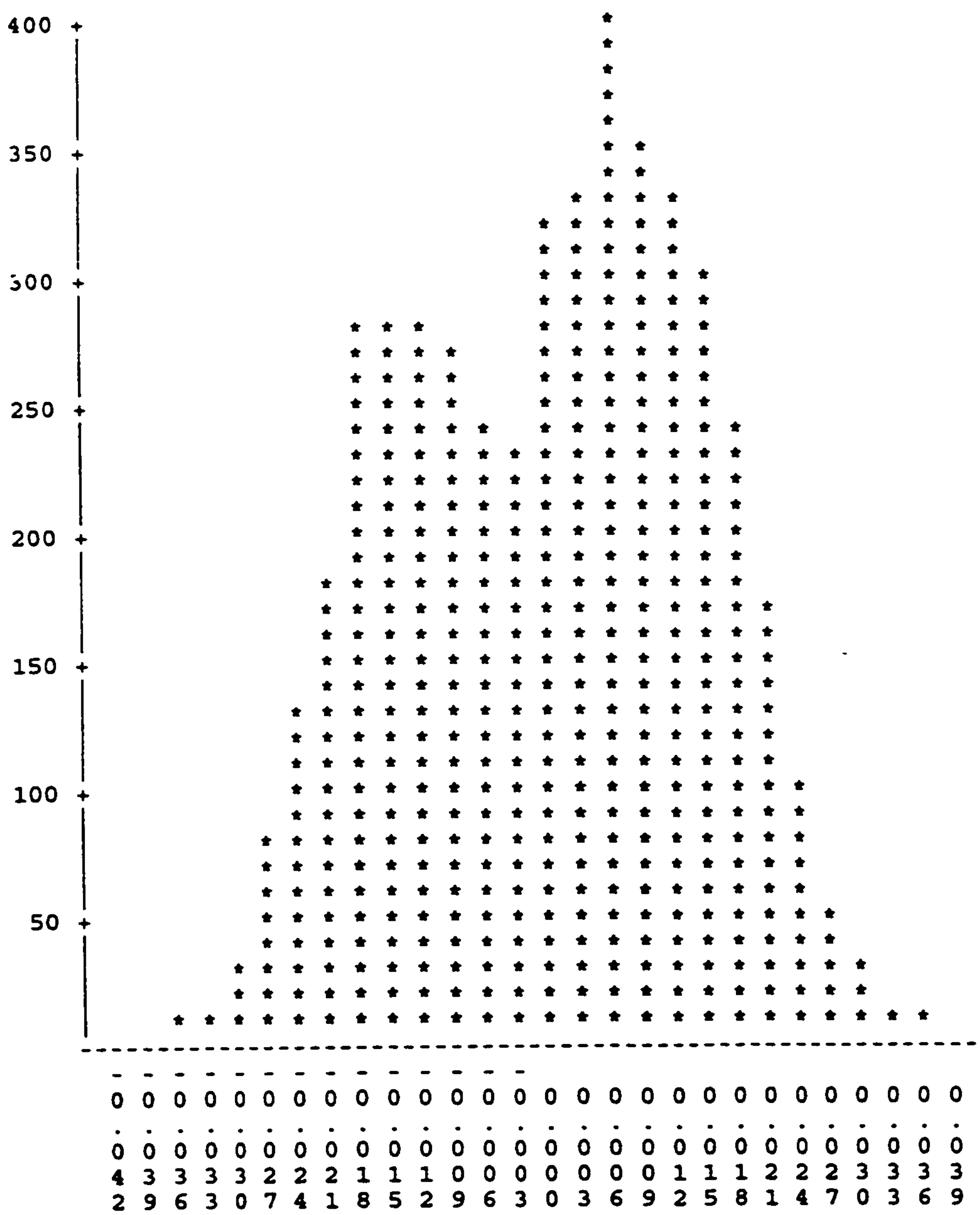
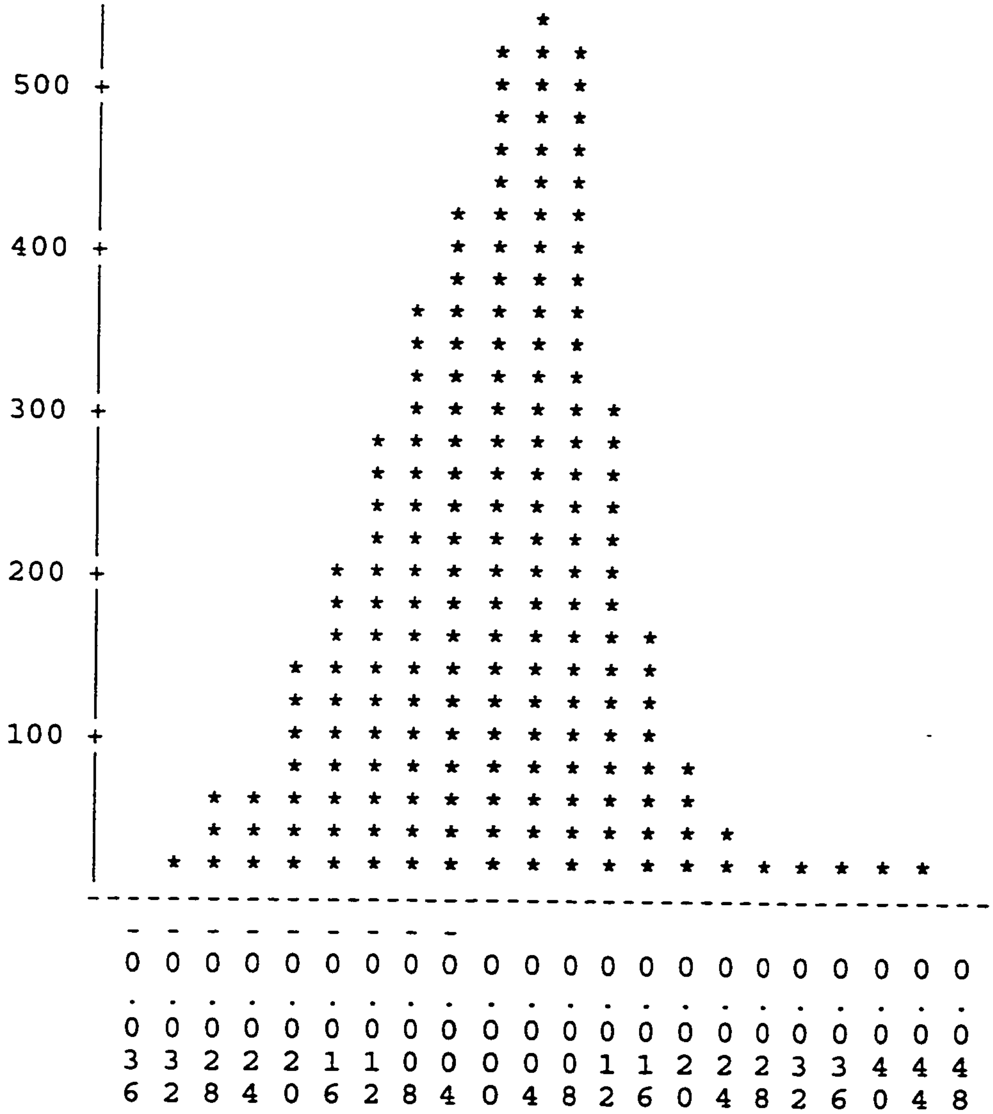


Diagram13:

This is the histogram generated from the bunting option.



10 Appendix C

Diagram14:

This is the residual plot generated from the batting normally option.

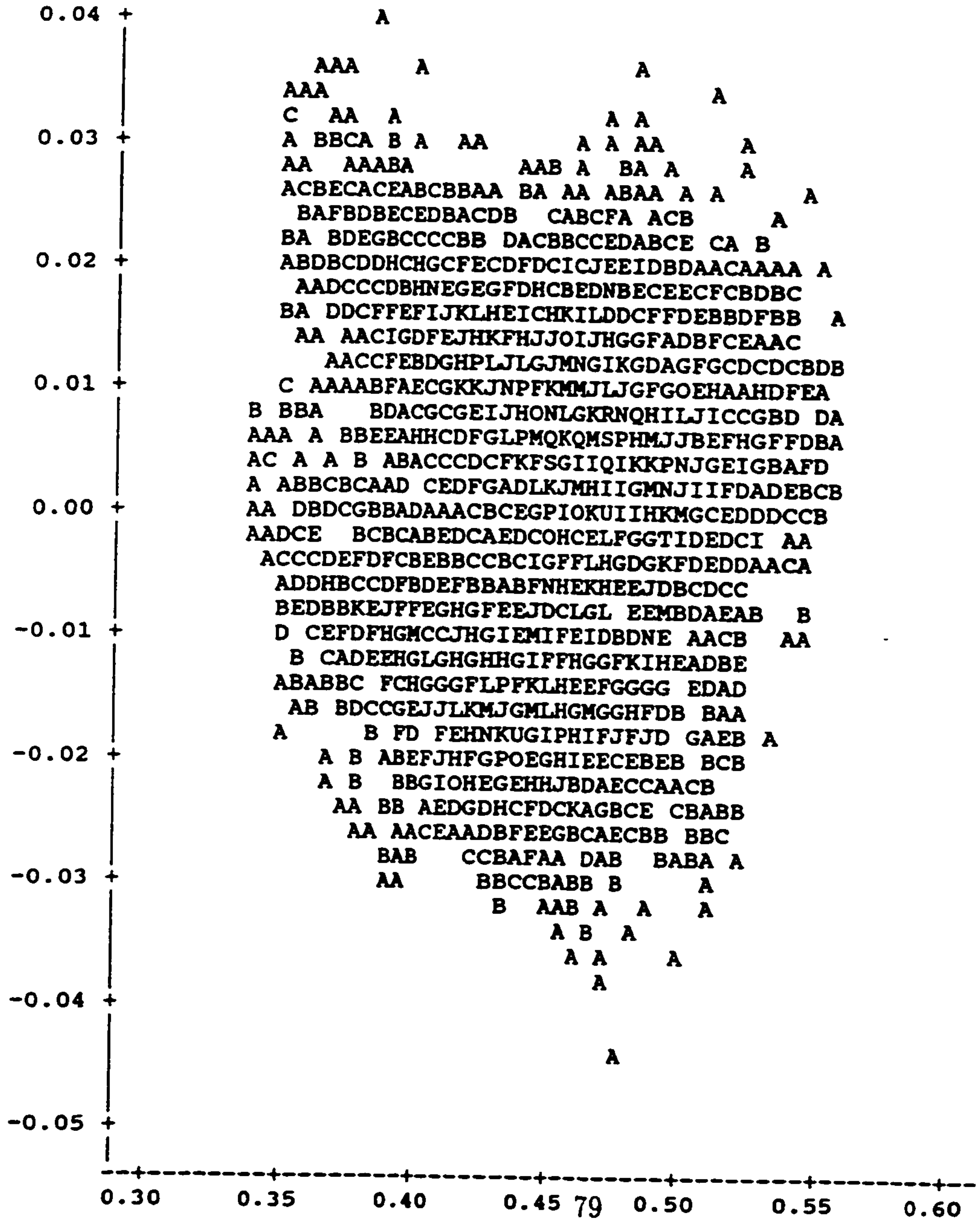
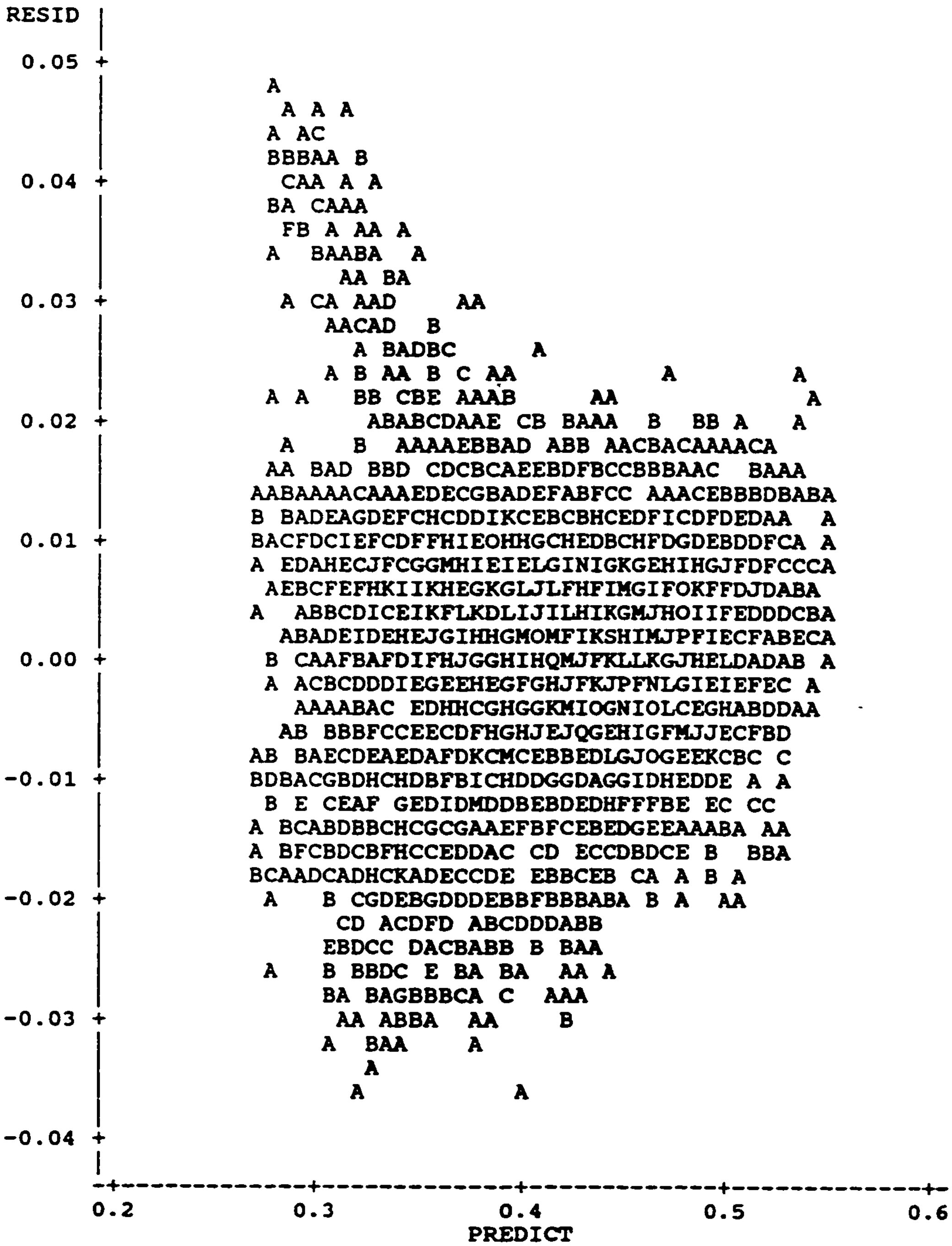


Diagram15:

This is the residual plot generated from the bunting option.



11 References

1. Christian Albright (1993),

   A Statistical Analysis of Hitting Streaks in Baseball,

   Journal of the American Statistical Association, Vol. 88, pp. 1175-1183.

2. Daniel Barry and D. A. Hartigan (1993),

   Choice Models for Predicting Divisional Winners in Major League Baseball,

   Journal of the American Statistical Association, Vol. 88, pp. 766-774.

3. Richard Bellman (1976),

   Dynamic Programming and Markovian Decision Process, With Particular Application to Baseball and Chess,

   Applied Combinatorial Mathematics, Chapter 7, pp. 221-236.

4. Richard Bellman (1977),

   Dynamic Programming and Markovian Decision Processes, With Application to Baseball,

   Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 77-85.

5. Alan C. Benander and Barbara A. Benander (1987),

N-team, M-loss Elimination Tournaments,

Journal of Recreational Mathematics, Vol. 19, pp. 266-267.

6. Jay Bennett (1993),

Did Shoeless Joe Jackson Throw the 1919 World Series?,

The American Statistician, Vol. 47, pp. 241-250.

7. Jay Bennett and D. A. Flueck (1984),

Player Game Percentage,

Proceedings of the Social Science Section, American Statistical Association, pp. 378-380.

8. Aaron Brown (1994),

The Probability of Breaking Sports Records,

Journal of Recreational Mathematics, Vol. 26, pp. 42-47.

9. William O. Cain (1977),

The Computer Assisted Heuristic Approach Used to Schedule the Major League Baseball Clubs,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science *and Systems*, pp. 32-41.

10. Bradley P. Carlin (1996),

    Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information,

    The American Statistician, Vol. 50, pp. 39-43.

11. Earnshaw Cook (1977),

    An Analysis of Baseball as a Game of Chance by the Monte Carlo Method,

    Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 50-54.

12. Earnshaw Cook (1966),

    Percentage Baseball

    Second Edition, M. I. T. Press, April 1966.

13. Andrew Cornish, Charles Fiedler, Brian S. Foss and Paul M. Sommers (1987),

    Eye-Pothesis Testing: Another Look at the 1986 World Series,

    Journal of Recreational Mathematics, Vol. 19, pp. 241-246.

14. D. A. D'Esopo and B. Lefkowitz (1977),

The Distribution of Runs in the Game of Baseball.

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 50-54.

15. Christopher T. Edwards (1996),

Double Elimination Tournaments: Counting and Calculating.

The American Statistician, Vol. 50, pp. 27-33.

16. Jill E. Erickson, Elizabeth A Loring and Lisa B. Seiden (1987),

The Boys of Summer of the Chis of Spring,

Journal of Recreational Mathematics, Vol. 19, pp. 161-163.

17. R. Allen Freeze (1977),

Monte Carlo Analysis of Baseball Batting Order,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 63-67.

18. Richard Friedlander, Stan Wagon and D. M. Bloom (1993),

Double Simpson's Paradox,

Mathematics Magazine, Vol. 66, pp. 268-269.

19. Yigal Gerchak (1994).

General Review of OR in Sports,

Handbooks in OR and MS, Vol. 6, Chapter 14.

20. Brett C. Holbrook , Kathryn L. Schenk and Neil C. Schwertman (1996).

More Probability Models for the NCAA Regional Basketball Tournament,

The American Statistician, Vol. 50, pp. 34-38.

21. Ronald A. Howard (1977),

Baseball a la Russe,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 89-93.

22. James L. Kepner (1985),

On the Probability that the Better Team Wins the World Series,

The College Mathematics Journal, Vol. 16, pp. 250-256.

23. James R. Lackritz (1990),

Salary Evaluation for Professional Baseball Players,

The American Statistician, Vol. 44, pp. 4-8.

24. G. R. Lindsey (1977),

A Scientific Approach to Strategy in Baseball,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 1-30.

25. Allen J. Michael and Carl R. Mitchell(1977).

The Valuation of a Baseball Player,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 42-45.

26. Pete Palmer (1977),

Batting Performance vs. Ball-Strike Count,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 31.

27. Arthur Peterson (1977),

Comparing the Run- Scoring Abilities of Two Different Batting Orders: Results of a Simulation,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 86-88.

28. Lawrence W. Robinson (1991).

Baseball Elimination's: An Application of Linear Programming,

Operations Research Letters, Vol. 10, pp. 67-74.

29. Edward Rozema (1994),

Round-Off, Batting Averages, and Ill-Conditioning.

The College Mathematics Journal, Vol. 25, pp. 314-317.

30. William Simon (1977),

Ball-to-the-Wall Effect: 1976 Perspective,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 46-47.

31. John H. Smith (1977),

Adjusting Baseball Standings for Strength of Teams Played,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 48-49.

32. Paul M. Sommers (1987),

Career Length and Playing Position in Major League Baseball: Any Connection?, *Journal of Recreational Mathematics, Vol. 19*, pp. 247-251.

33. Paul M. Sommers (1989),

The Golden Boys of Summers, 1987,

Journal of Recreational Mathematics, Vol. 21, pp. 161-165.

34. Paul M. Sommers (1994),

Probably the Greatest Hitter Who Ever Lived,

Journal of Recreational Mathematics, Vol. 26, pp. 32-35.

35. STATS INC. (1994)

STATS Player Profiles 1995

First Edition, STATS Publishing, November 1994.

36. STATS INC. (1995)

STATS Player Profiles 1996

First Edition, STATS Publishing, November 1995.

37. Richard E. Trueman (1976),

A Computer Simulation Model of Baseball: With Particular Application to Strategy Analysis,

Management Science in Sports, Vol. 4, Studies in the Management Sci-

ences, pp. 1-15.

38. Richard E. Trueman (1977),

Analysis of Baseball as a Markov Process,

Optimal Strategies in Sports, Vol. 5, Studies in Management Science and Systems, pp. 68-76.

12 Vita Auctoris

Mathew Chandler was born in 1968 in Ottawa, Ontario. He graduated from Kennedy High School in Windsor in 1987. From there, he went to the University of Windsor where he earned a B. A. in Psychology in 1990 and a B. Sc. in Honours Mathematics in 1996. He is currently a candidate for a Master's degree in Statistics at the University of Windsor and hopes to graduate in Fall 1997.

IMAGE EVALUATION TEST TARGET (QA-3)

150mm

6"

APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989